Técnicas estadísticas multivariantes para la caracterización de la Apolipoproteína E4

Realizado por:
Abel García Rubio
Directores:
José Miguel Arbonés Mainar
Ana Pérez Palomares
Beatriz Lacruz Casaucau

Trabajo de Fin de Máster en Modelización e investigación matemática, estadística y computación

Resumen

La Apolipoproteína E (ApoE) es un gen polimórfico que desempeña una importante función en la regulación del metabolismo de colesterol, ácidos grasos y la homeostasis de la glucosa. Este gen ApoE posee tres alelos (variantes) principales ApoE2, ApoE3 y ApoE4. Cada individuo se clasifica en una de las tres variantes y el interés de su estudio radica en que la presencia de un alelo u otro tiene consecuencias fisiológicas profundas en el ser humano. El objetivo del estudio consiste en emplear técnicas multivariantes que permitan determinar qué variables se encuentran asociadas con la variante ApoE4, que es un gen que puede guardar relación con enfermedades cardíacas o alzheimer. En este estudio utilizamos datos clínicos, biométricos y muestras biológicas de trabajadores sanos voluntarios de la factoría de automóviles de General Motors España (GME) en Figueruelas (Zaragoza).

Se utilizarán técnicas como regresión logística, regresión logística penalizada, árboles de clasificación y particionamiento recursivo basado en modelos. Veremos qué variables como Proteína C Reactiva, Lipoproteína o incluso el cociente entre Apolipoproteína B y Apolipoproteína A1 revelan más información a la hora de determinar la pertenencia de la observación al grupo ApoE4.

Abstract

Apolipoprotein E (ApoE) is a polymorphic gene which perfoms an important role in the regulation of cholesterol metabolism, fatty acids and glucose homeostasis. ApoE gene has three alleles (variants) ApoE2, ApoE3 and ApoE4. A individual is classified into one of the three variants and the interest of their study is that the appearance of one allele or another has deep physiological consequences in humans. The aim of the study is to use multivariate techniques which allow us to determine what variables are associated with the ApoE4 variant, which is a gene that may be related to heart or Alzheimer's disease. In this study, we used clinical, biometric and biological samples of healthy volunteer workers from the General Motors España (GME) automobile factory in Figueruelas (Zaragoza).

Techniques like logistic regression, penalized regression, classification trees and recursive partitioning based on models are used. We will see what variables such as ProteinaCReactiva, Lipoprotein or even the quotient between Apolipoprotein B and Apolipoprotein A1 give relevant information when determining the membership of the ApoE4 variant.

Agradecimientos

Me gustaría agradecer a Ana Pérez, Beatriz Lacruz y José Miguel Arbonés Mainar por su infinita paciencia, dedicación y sobretodo por ayudarme en la realización de este Trabajo Fin de Máster.

Gracias. :)

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Planteamiento del problema

1.1. Introducción

El gen de la Apolipoproteína E es un gen polimórfico y tiene tres alelos principales: *ApoE2*, *ApoE3* y *ApoE4* que se definen a partir de los polimorfismos de nucleótido simple (SNP, del inglés *Single Nucleotide Polymorphims*): SNP112 y SNP158. *ApoE3* es el alelo más común, ya que está presente en el 70-80 % de los individuos; *ApoE4* está en el 15 % de la población; mientras que *ApoE2* representa solo alrededor del 5 %. La presencia de *ApoE4* es un factor de riesgo para enfermedades cardiovasculares y el Alzheimer, mientras que *ApoE2* está asociado a una disminución de dicho riesgo [4]. La *ApoE* tiene un papel fundamental en el metabolismo de los lípidos y en la actividad de las células adiposas, por lo que resulta de gran interés investigar la relación existente entre los factores relacionados con la obesidad y las enfermedades cardiovasculares asociadas y los distintos alelos del gen [1]. *ApoE4* también puede contribuir en el desarrollo del síndrome metabólico[1] [2]. Tanto en [1] como en [2] se citan otros estudios encaminados al estudio de este tipo de relaciones. El objetivo general de este Trabajo de Fin de Máster se sitúa en esta misma línea y pretende determinar qué factores asociados a una determinada enfermedad están asociados con los distintos alelos del gen *ApoE* y en qué sentido, ya que así el propio gen podría ayudar a identificar sujetos con riesgo de padecer ciertas enfermedades, lo que ayudaría a mejorar la acción preventiva.

Para este trabajo disponemos de datos recogidos en 2010 a trabajadores de General Motors España (fábrica de automóviles situada en Figueruelas (Zaragoza)) a los que no se les había diagnosticado nunca una enfermedad cardiovascular. Estos datos provienen del proyecto de investigación AWHS (*Aragón Workers Health Study*) que comenzó en 2009 fruto de la colaboración entre el Gobierno de Aragón con el Instituto Aragonés de Ciencias de la Salud y el Centro Nacional de Investigaciones Cardiovasculares (CNIC). Su objetivo es avanzar en el conocimiento y la prevención de las enfermedades cardiovasculares e intentar determinar las causas por las que unas personas desarrollan más frecuentemente que otras infarto de miocardio e ictus. Para el estudio se cuenta con más de 5.000 voluntarios, todos ellos trabajadores de la fábrica antes mencionada, los cuales son sometidos anualmente a un análisis de sangre, suero, plasma y orina, y se realiza un reconocimiento físico en el que se recoge: peso, altura, perímetro de cintura abdominal, presión arterial, exploración física y electrocardiograma. Este estudio es el mayor de estas características hecho en España y uno de los más importantes de Europa.

En este Trabajo de Fin de Máster nos centraremos en el conjunto de trabajadores varones (puesto que el número de mujeres que participan en el estudio es muy pequeño) y aplicaremos técnicas de clasificación que nos permitan determinar qué variables están más asociadas con el genotipo E4 [2]. Comenzaremos construyendo modelos de regresión logística (con selección de variables hacia adelante) y

[1] El síndrome metabólico está definido como un conjunto de factores de riesgo que incluyen la obesidad, la presión sanguínea alta, la hiperglucemia (aumento anormal de la cantidad de glucosa en sangre) y la dislipidemia (colesterol HDL bajo o triglicéridos altos, en sangre).

[2] Inicialmente se realizó un estudio descriptivo con los tres grupos de *ApoE* (que se incluye en la sección 2.1) como exploración previa de los datos, pero finalmente se decidió por recomendación de los expertos centrar el estudio en la caracterización del genotipo E4.

regresión logística penalizada que son técnicas que permitirán determinar los factores que más influyen en el genotipo E4. También se construirán árboles de clasificación que además de seleccionar variables influyentes proporciona una forma sencilla de encontrar interacciones entre dichos factores. Estas técnicas se complementarán con la de particionamiento recursivo para construir modelos segmentados. Esta técnica extiende los modelos de regresión logística y los árboles de clasificación, permitiendo refinar los modelos anteriores. Sabemos de antemano, por la naturaleza del problema, que es difícil que los métodos de clasificación proporcionen modelos con buena capacidad predictiva, pero el interés no recae en ningún caso en clasificar nuevos individuos, sino en tratar de descubrir qué variables los diferencian y con qué enfermedades están relacionadas, así como averiguar el papel de estas variables en esta diferenciación. Es por ello por lo que se han elegido métodos de clasificación que proporcionen modelos fácilmente interpretables.

El análisis de los datos se ha realizado con el programa estadístico R (1.0.1) [3]. Los paquetes utilizados para cada una de las técnicas se citarán en el correspondiente apartado. La memoria de este Trabajo de Fin de Máster se organiza como sigue: este capítulo continúa con la descripción de las bases de datos disponibles para la realización del estudio y la descripción de cada una de las variables contenidas en las mismas. En el capítulo 2 se presentan los fundamentos teóricos de cada una de las técnicas anteriormente mencionadas. Además, se explica la técnica de validación cruzada que se utilizará para la selección de los modelos y las medidas de bondad del ajuste utilizadas. El capítulo 3 está dedicado al preprocesamiento de los datos y su análisis exploratorio. En el capítulo 4 se incluyen los modelos construidos y se finaliza con una sección dedicada a las conclusiones obtenidas.

1.2. Bases de datos

La primera base de datos recibe el nombre de 'Síndrome Metabólico' en donde se puede encontrar la siguiente información sobre el paciente:

- SEQN: variable de identificación de los sujetos (también se encuentra en 'GM2010 variables' permitiendo realizar una unión entre las dos bases con una cardinalidad de interrelación, a priori, de 1 a 1).
- metSindrom: variable dicotómica sobre si el paciente tiene síndrome metabólico (0=NO / 1=SI).
- APOE: clasificación de la *ApoE* de cada individuo en una de las 3 categorías (E2, E3 y E4).
- SNP112 y SNP158: cada uno de los polimorfismos a partir de los cuales se define la *ApoE*.

La segunda base de datos recibe el nombre de 'GM2010 variables' en donde es posible encontrar información sobre las analíticas y el CRD (Cuaderno de Recogida de Datos que contiene otro tipo de características como el peso, altura,...) que se han realizado al paciente.

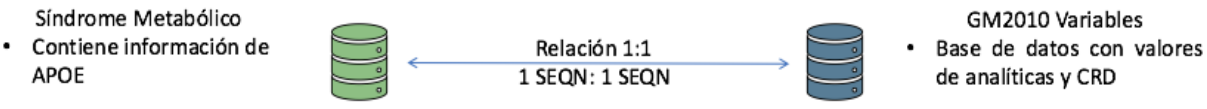

Figura 1.1: Relación entre las bases de datos.

El primer paso de todos es combinar las dos bases para identificar correctamente a los individuos. En el proceso se retiran columnas que aportan poca información ya sea porque tienen valores constantes o porque es un código interno carente de significado en el estudio.

1.3. Descripción de las variables

Las variables presentes en esta base de datos tienen una interpretación médica y que sin un estudio previo de cada una de ellas sería un gran obstáculo comprender este trabajo. Por tanto, en esta sección intentaremos dar una breve explicación de cada una de ellas obtenida de las referencias [4], [5] y [6]:

- **Bilirrubina**

  Es un pigmento amarillo que se encuentra en el hígado y se produce tras la eliminación de los glóbulos rojos viejos. Los valores fuera de los parámetros normales pueden deberse a enfermedades como hemólisis o hepatitis.

  Los parámetros normales considerados van desde 0.1 hasta 1.2 mg/dL (miligramos por decilitro).

- **Calcio**

  Todas las células necesitan calcio para trabajar, además es importante para la función cardíaca y ayuda con la contracción muscular, las señales nerviosas y la coagulación sanguínea. Los valores fuera de lo común se encuentran relacionados con calambres musculares o incluso insomnio.

  Los valores normales van de 8.5 a 10.2 mg/dL.

- **Colesterol**

  El colesterol es unos de los principales elementos para que las células puedan operar correctamente, pero a su vez presenta un gran inconveniente, si se encuentra en exceso puede tapar las arterias y provocar enfermedades cardíacas.

  Un colesterol total de 120 a 220 mg/dL o menos se considera ideal.

- **HDL Colesterol**

  HDL (*High-Density Lipoproteins*) significa lipoproteínas de alta densidad. Se le llama colesterol 'bueno' porque transporta el colesterol de otras partes del cuerpo al hígado para ser eliminado.

  Los niveles deseados de HDL colesterol están entre los 40 y los 60 mg/dL.

- **Creatinina y Creatinina en orina**

  La creatinina es un producto orgánico de deshecho que produce el cuerpo y que es filtrado por los riñones. Generalmente este valor se obtiene para determinar si los riñones funcionan correctamente, ya que un mal funcionamiento puede ocasionar graves problemas.

  Un resultado normal en sangre es de 0.7 a 1.3 mg/dL para los hombres.

  Un resultado normal en orina es >100mg/dL.

- **GGT**

  La Gamma-Glutamil Transferasa, GGT, es una enzima que se encuentra en el hígado y riñones en niveles elevados. Esta medida se utiliza para detectar enfermedades en el hígado o vías biliares. Los valores anormales en este parámetro pueden desencadenar enfermedades como: diabetes, enfermedades pulmonares, enfermedades en el páncreas...

  El rango normal para adultos varía entre 0 hasta 55 U/L (unidades por litro).

- **Glucosa**

  La glucosa es una fuente importante de energía para la mayoría de las células del cuerpo, incluyendo las del cerebro. Principalmente se lleva un estricto control de este parámetro en pacientes con diabetes.

  Los valores basales de glucemia normales oscilan entre 70 y 100 mg/dl.

- **Enzimas AST/GOT**

  La enzima ASpartato Aminotransferasa es un signo de una enfermedad hepática. Una de las razones en las que este parámetro presenta valores fuera de lo común es debido a un ataque cardíaco.

  El rango normal es de 0 a 50 U/L.

- **Encimas ALT/GPT**

  La enzima ALanina Transaminasa se encuentra en grandes cantidades en el hígado. Una lesión en el hígado provoca la liberación de ALT en la sangre. También se puede utilizar para determinar si la persona tiene daño hepático (son los diversos factores que causan daño en el hígado).

  El rango normal en los hombres es de 0 a 50 U/L.

- **Triglicéridos**

  Los triglicéridos son un tipo de grasa que se almacena en el cuerpo cada vez que consumes más calorías de las necesarias. Una diabetes mal controlada puede influir en este parámetro.

  Los valores de esta variables se agrupan de la siguiente manera:

  - Normal: menos de 150 mg/dL.
  - Limite: 150 a 199 mg/dL.
  - Alto: >200 mg/L.

- **Ácido úrico**

  El ácido úrico es un compuesto químico que se crea cuando el cuerpo descompone sustancias llamadas purinas. Las purinas son generadas por el cuerpo y que además se pueden encontrar en algunos alimentos y bebidas. Los valores fuera de lo común pueden indicar que el paciente presenta diabetes de tipo 2, esclerosis múltiple o enfermedades cardiovasculares.

  Los valores normales están entre 3.5 y 7.2 miligramos por decilitro (mg/dL).

- **Apolipoproteína A1**

  La Apolipoproteína A-1 (Apo A1) pertenece al grupo de las proteínas de alta densidad (HDL). Su papel metabólico es la activación de la lecitina colesterol aciltransferasa (HMG Co) que cataliza la esterificación del colesterol. Una vez esterificado el colesterol puede ser transportado hacia el hígado donde se metaboliza y es excretado. Esta apolipoproteína se encuentra relacionada con HDL colesterol. Los valores bajos indican el posible desarrollo de enfermedades como arteriosclerosis, enfermedades circulatorias, coronarias, etc.

  Los valores normales están entre 90 y 170 mg/dL.

- **Apolipoproteína B**

  La Apolipoproteína B-100, también conocida como apolipoproteína B o Apo B, es una proteína implicada en el metabolismo de los lípidos, siendo el principal constituyente proteico de lipoproteínas de muy baja densidad (VLDL) y de baja densidad (LDL, o colesterol "malo"). Los valores altos implican el desarrollo de enfermedades ya mencionadas en la variable anterior.

  Su rango de valores se encuentra entre 56 y 162 mg/dL.

- **Lipoproteína(a)**

  Las lipoproteínas son moléculas hechas de proteínas y grasa. Transportan el colesterol y sustancias similares a través de la sangre. Un valor alto puede ser un factor de riesgo de cardiopatía.

  Los valores normales están por debajo de 30 mg/dL.

- **Proteína C Reactiva**

  La proteína C reactiva (PCR) es producida por el hígado. El nivel de PCR se eleva cuando hay inflamación en todo el cuerpo. Pertenece a un grupo de proteínas llamadas 'reaccionantes de fase aguda' que aumentan en respuesta a la inflamación. Los valores anormales en este parámetro pueden ser signos de padecer enfermedades cardiovasculares.

  Su rango de valores normales es 0-0.5 mg/dl.

- **TSH**

  La hormona estimulante de la tiroides, o TSH por sus siglas en inglés, es producida por la hipófisis y se encarga de ordenar a la glándula tiroides producir y secretar las hormonas tiroideas en la sangre.

  Los valores normales pueden fluctuar de 0.4 a 4.0 mlU/L.

- **T4 libre**

  La tiroxina es una hormona producida por la glándula tiroides que ayuda a controlar el metabolismo además del crecimiento.

  Un rango normal y típico va de 1.2 y 2.2 ng/dL.

- **Microalbúmina**

  Es una proteína que se encuentra en los riñones y sirve para detectar si el paciente puede desarrollar una enfermedad del riñón, o por ejemplo enfermedades como diabetes o hipertensión.

  Los niveles normales de albúmina en la orina son de menos de 30 mg/dL.

- **Hemoglobina Glicosilada**

  La hemoglobina glicosilada mide el número de glóbulos rojos que se encuentran ligados a una molécula de glucosa.

  Los parámetros valores de la hemoglobina los podemos clasificar en el siguiente rango; < 5.7 normal, 5.7 - 6.4 riesgo de diabetes y >= 6.5 diabetes.

- **Presión diastólica y sistólica**

  La presión sistólica es la presión máxima que se alcanza en el sístole (contracción del corazón) y la presión diastólica es la mínima presión de la sangre contra las arterias y ocurre durante el diástole (relajación del corazón). Una presión sistólica elevada puede aumentar el riesgo de padecer enfermedades coronarias, o por el contrario, si la presión arterial es baja puede indicar que hay problemas con el corazón.

  Los valores normales para presión sistólica se encuentran entre 100 y 140 mm de Hg y para presión diastólica entre 60 y 90 mm de Hg.

- **Medicamentos**

  En este caso son variables dicotómicas que indican si el paciente se medica por padecer diabetes, dislipemia, hipertensión o formación de trombos, siendo la unidad en la situación favorable. Estas variables pueden tener influencia con las variables cuantitativas como por ejemplo, el caso de padecer diabetes, los resultados de glucosa pueden salir normales si la persona se encuentra en tratamiento.

- **Otras variables**

  También hay variables adicionales como altura, peso, perímetro abdominal y consumo de alcohol a la semana.

Finalmente tenemos en este conjunto de datos un total de 37 variables.

Capítulo 2

Metodología

"I keep six honest serving-men
(They taught me all I knew)
Their names are What and Why and When
And How and Where and Who."

Rudyard Kipling[1]

2.1. Introducción

El amplio campo de las matemáticas ofrece diversas técnicas para crear, mejorar e incluso validar los modelos. Las técnicas más relevantes para este trabajo son recogidas en este capítulo proporcionando al lector un breve resumen de cómo funcionan y con qué finalidad se pueden emplear.

En primer lugar el lector podrá informarse sobre la regresión logística además de hacerse una idea de por qué ha sido elegida esa técnica. La siguiente sección es dedicada a la regresión penalizada, técnica hermana de la regresión logística, donde son explicadas las diferentes técnicas de penalización. Posteriormente se pasará a los árboles de decisión obteniendo así una descripción concisa que permitirá conocer cómo opera el árbol frente a los valores perdidos. Esto es ventaja comparado con la regresión logística ya que esta última no trabaja con valores perdidos. A continuación podrá familizarse con el particionamiento recursivo basado en modelos en donde son combinadas las técnicas de regresión logística y los árboles de clasificación. Para finalizar el capítulo se concluirá con las técnicas de validación cruzada (procedimiento que ayuda a mejorar la estimación de los parámetros obtenidos) y la curva ROC (que nos permite determinar la bondad del modelo).

2.2. Modelos de regresión logística

Un modelo de regresión logística es un análisis de regresión que establece una relación entre una variable categórica (variable respuesta o dependiente) y unas variables explicativas (predictores o independientes). Los modelos cumplen tres finalidades, [12]: clarificar, cuantificar y clasificar. Clarificar la existencia de interacción y confusión entre covariables respecto a la variable dependiente; cuantificar la importancia de la relación entre cada una de las variables independientes y la variable dependiente, y clasificar individuos dentro de las categorías de la variable dependiente. En este estudio nos centraremos en los dos primeros objetivos, es decir, detectaremos el conjunto de variables (junto con posibles interacciones) asociadas a la apolipoproteína E4.

Partiremos de una variable respuesta dicotómica con valores 0 y 1, en nuestro caso, $Y = 1$ es pertenecer a E4 e $Y = 0$ en otro caso. En un modelo logístico binario, la esperanza de la variable respuesta

[1] Cita obtenida del libro 'How to stop worring and start living' de Dale Carnegie, pág 53.

(Y), es decir, la probabilidad de que $Y = 1$, dada la información de las variables predictoras, se modela de la siguiente manera:

$$P(Y = 1/\vec{x}) = \frac{exp(\vec{x}'\vec{\beta})}{1 + exp(\vec{x}'\vec{\beta})} \tag{2.1}$$

siendo $\vec{x} = (x_1, x_2, x_3, \ldots, x_P)'$ los valores de las variables explicativas y $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)$ el vector de coeficientes. La regresión logística también se puede utilizar como método para la estimación de la razón de disparidad[2] (odds):

$$odd(\vec{x}) = \frac{P(Y = 1/\vec{x})}{1 - P(Y = 1/\vec{x})} = exp(\vec{x}'\vec{\beta})$$

adoptando su logaritmo una expresión lineal en $\vec{\beta}$, lo que guarda cierta semejanza con la regresión lineal.

$$Log(odd(\vec{x})) = In\left(\frac{P(Y = 1/\vec{x})}{1 - P(Y = 1/\vec{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Luego los coeficientes de β_i muestran el efecto que tiene la variable x_i sobre la variable dependiente Y. Así, un coeficiente de β cercano a cero, o en su defecto un odds-ratio cercano a la unidad, indican que la variable x_i no tiene efecto alguno sobre Y.

2.2.1. Métodos de estimación de parámetros

A diferencia de la regresión lineal, que utiliza los mínimos cuadrados, en la regresión logística se emplea el método de Máxima Verosimilitud (MV) para llevar a cabo la estimación de los parámetros del modelo.

Supondremos una muestra aleatoria (x_i, y_i), $i = 1, 2, ..., n$ y denotemos por $p_i = P(Y = 1/\vec{x_i})$. La función de probabilidad para una respuesta y_i cualquiera es:

$$F(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \qquad y_i = 0, 1$$

y para la muestra:

$$F(y_1, \ldots, y_n) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}.$$

Tomando logaritmos llegamos a la siguiente función:

$$logF(\beta) = \sum_{i=1}^{n}(y_i log p_i + (1 - y_i) log(1 - p_i)) \tag{2.2}$$

donde la dependencia de β se obtiene al sustituir en 2.1 en esta expresión. Por tanto los parámetros β del modelo se obtienen realizando la derivada correspondiente a la función 2.2 e igualando a cero. De modo que tendremos $p + 1$ ecuaciones no lineales que se resuelven mediante métodos numéricos.

2.2.2. Criterios para incluir o no un parámetro

Existen distintas formas de evaluar la inclusión de un parámetro en el modelo entre las que destacaremos AIC ('Akaike Information Criterion'), BIC ('Bayesian Information Criterion') o criterio de Mallows. El criterio utilizado en estas regresiones logísticas es el AIC que mide la bondad de ajuste a partir de la máxima verosimilitud del modelo, penalizada con la complejidad del modelo (número de parámetros).

[2]Cociente entre ocurrencia de un evento frente a la no ocurrencia del mismo.

2.2.3. Métodos de selección de variables

Una aproximación ingenua al problema consistiría en estudiar la reducción de una cierta medida, por ejemplo BIC o AIC, originada por la incorporación de una nueva variable al modelo. Desafortunadamente este procedimiento no tiene en cuenta que la reducción del criterio fijado por la inclusión de una nueva variable puede estar regida por las ya existentes en la ecuación ajustada.

Por consiguiente una posible solución sería, dados p regresores, evaluar los subconjuntos de regresores existentes y efectuar las regresiones, reteniendo aquélla que mejor se ajuste al criterio de bondad de ajuste seleccionado. El inconveniente recae en que es preciso realizar un gran volumen de cálculo. Una manera de esquivar este obstáculo sería utilizar la regresión escalonada ('stepwise regression') que proporciona modelos que son óptimos o muy próximos al óptimo. Hablaremos de los procedimientos de regresión escalonada 'hacia adelante' ('forward selection procedure'), la regresión escalonada 'hacia atrás' ('backward elimination') o regresión escalonada mixta.

En cada instante en que la regresión 'hacia adelante' es ejecutada, tendremos una ecuación de regresión provisional que incluye algunas variables (regresores) y no otras (auxiliares). Al inicio del procedimiento, la ecuación de apoyo no incluye ninguna variable, y después sigue el siguiente esquema (ver referencia [11]):

- **Paso 1:** Calcula los estadísticos asociados Q_h para todos los regresores ausentes.
- **Paso 2:** Sea Q_h^* el máximo estadístico de los calculados, entonces: si $Q_h^* < U$, siendo U un umbral prefijado, finaliza; la ecuación provisional es la definitiva. Si, por el contrario, $Q_h^* \geq U$, se introduce la variable en la ecuación de regresión.
- **Paso 3:** Si no quedan regresores ausentes, finalizar procedimiento. En caso contrario, reiniciar los calculos en **paso 1**.

En resumen, se trata de incluir variables de una en una, por orden de mayor contribución al disminuir BIC o AIC. La regresión paso a paso 'hacia adelante' es una excelente opción para resolver problemas de multicolinealidad y evitar que variables redundantes entren en el modelo.

El procedimiento de regresión 'hacia atrás' procede de manera similar pero comenzando con una ecuación que incluye todos los regresores. Por el contrario, el procedimiento mixto, alterna la inclusión y exclusión de las variables en el modelo, permitiendo la exclusión de variables que ya habían sido consideradas anteriormente. Los criterios de entrada y salida se fijan especificando sendos valores $U_{entrada}$ y U_{salida} que deben ser no alcanzados (superados) por Q_h^* correspondiente para que la variable pueda ser incluida (o excluida).

2.3. Regresión penalizada

La regresión penalizada es un método de regresión empleado en situaciones en las que puede existir alta correlación entre las variables. El objetivo es mejorar la predicción e interpretabilidad de los modelos que se obtienen maximizando la función de verosimilitud. Los métodos de penalización se basan en la contracción de los parámetros imponiendo que la norma de estos no supere un determinado valor.

Dentro de las técnicas de penalización se encuentran la regresión ridge (penaliza con la norma L_2 sobre los coeficientes), la regresión lasso (penaliza con la normal L_1) y una combinación de ambas conocida como Elastic net, [13]. Así el método de penalización consiste en:

$$min\{logF(\beta)\} = \sum_{i=1}^{n}(y_i log p_i + (1-y_i)log(1-p_i)) \tag{2.3}$$

sujeto a $\alpha||\beta||_1 + (1-\alpha)||\beta||_2 <= t$, donde α en $[0,1]$ y t son parámetros a determinar. Los casos $\alpha = 0$ y $\alpha = 1$ corresponden a la regresión ridge y lasso respectivamente.

En nuestro caso nos centraremos en la regresión lasso y elastic net puesto que proporcionan modelos parsimoniosos; en definitiva son métodos para la selección de variables. Para utilizar estas funciones utilizaremos el paquete *glmnet* de R [14]

2.4. Árboles de decisión

En la referencia [7] se puede encontrar que define el método de particionamiento como un problema de clasificación de n observaciones en C clases. El modelo creado se encarga de dividir las observaciones en k grupos terminales asignados a una determinada clase utilizando cualquier tipo de datos (numéricos o categóricos).

Para describir a los árboles es necesario conocer los siguientes términos:

- π_i: $i = 1,2,3\ldots,C$, es la probabilidad a priori de cada clase.
- $L(i,j)$: $i = 1,2,3\ldots,C$ matriz de pérdidas para la incorrecta clasificación con la diagonal nula, $L(i,i) = 0,\ \forall i, i = 1,2$.
- **Split:** Es cada una de las divisiones del conjunto de datos. Para formar un *split* se selecciona la variable y el valor de la misma que 'mejor' divide al conjunto de datos en dos subconjuntos. Por ello se utilizan distintos criterios como el Indice Gini y la ganancia de información de los que hablaremos más adelante.
- **Nodo:** Cada variable seleccionada para formar una división (*split*) será un *nodo* del árbol y cada una de las divisiones será una *rama*. La primera variable seleccionada conforma el *nodo raíz* y un nodo que ya no se subdivide es un *nodo terminal*. Un nodo terminal tiene asignada una distribución de probabilidad de forma que un nuevo individuo que ha llegado hasta allí se clasifica en la clase de mayor probabilidad según dicha distribución. Denotaremos esta variable como A. También denotaremos con los subindices L o R al nodo A para referirnos a los hijos del nodo y distinguiendo entre aquellos que cumplen la desigualdad de segmentación (nodo izquierdo, A_L) y los que la incumplen (nodo derecho, A_R).
- **Profundidad de un nodo:** Es la longitud del único camino existente desde el nodo raíz hasta el dicho nodo.
- **Nivel:** Conjunto de nodos con igual profundidad.
- $\tau(x)$**:** clase de la observación x, donde x es un vector de variables predictoras.
- $\tau(A)$**:** clase asignada al nodo A.
- n_i**,** n_A**:** número de observaciones en muestra que tiene clase i y número de observaciones en el nodo A. Por tanto, n_{iA} es el número de muestras de la clase i en el nodo A.
- $P(A)$**:** probabilidad del nodo A (para futuras observaciones) $= \sum_{i=1}^{C} \pi_i P\{x \in A | \tau(x) = i\} \approx \sum_{i=1}^{C} \pi_i n_{iA}/n_i$.
- $p(i|A)$**:** probabilidad de que la clase i se encuentre en el nodo A. $= \pi_i P\{x \in A|\tau(x) = i\}/P(x \in A) \approx \pi_i(n_{iA}/n_i)/\sum \pi_i(n_{iA}/n_i)$.
- $R(A)$**:** Riesgo del nodo A, Riesgo de $A = \sum_{i=1}^{C} p(i|A)L(i,\tau(A))$ donde $\tau(A)$ es elegido para minimizar el riesgo.

  Denotaremos T al árbol de clasificación, entonces:

- $R(T)$**:** Riesgo del modelo (o árbol) T, $= \sum_{j=1}^{k} P(A_j)R(A_j)$ donde A_j son los nodos terminales del árbol.

La figura 2.1 muestra un ejemplo de un árbol con 5 nodos y tres niveles de profundidad.

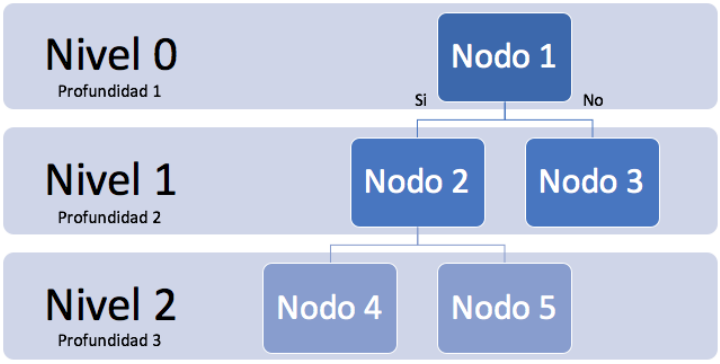

Figura 2.1: Ejemplo de la estructura de un árbol de decisión.

2.4.1. Criterios de segmentación

Un árbol de clasificación se construye siguiendo este esquema: se busca la variable que mejor divida el conjunto de datos según un criterio fijado de antemano; una vez que los datos están segmentados, se vuelve a aplicar el proceso a cada subgrupo; y de forma recursiva se continúa aplicando a los subgrupos restantes hasta satisfacer un criterio de parada. En cada nodo A que se divide en dos hijos A_L y A_R se cumple la siguiente desigualdad:

$$P(A_L)R(A_L) + P(A_R)R(A_R) \leq P(A)R(A) \tag{2.4}$$

Una manera de construir un árbol sería eligiendo la mejor división que maximice el riesgo, ΔR, pero no resulta lo más conveniente. Como alternativa, en la referencia [7] proponen las siguientes funciones de impureza que son reemplazadas el riesgo del nodo en la desigualdad 2.4.

- Criterio de Gini:

  Es utilizado por los árboles de clasificación, o regresión, como medida para determinar la pureza de los nodos. Se basa en calcular en cada nodo la probabilidad de asignar mal un individuo que ha sido seleccionado aleatoriamente según la distribución de probabilidad subyacente en ese nodo y se vuelve a clasificar según esa misma distribución de probabilidad, es decir,

  $$I_G(A) = \sum_{i=1}^{K} p_i(1-p_i) = 1 - \sum_{i=1}^{K} p_i^2 \tag{2.5}$$

  donde K es el número de clases y p_i es la probabilidad de la clase i.

- Criterio de información:

  El criterio de información, o ganancia de información, se encarga de medir la entropía, de modo que se elegirán los hijos que proporcionen una mayor ganancia de información. Se calcula como:

  $$I_I(A) = \sum_{i=1}^{K} p_i ln p_i$$

2.4.2. Criterios de parada

Todos los árboles de clasificación necesitan un criterio de parada para evitar que el crecimiento del mismo se extienda hasta que cada caso ocupe su propio nodo. Estableciendo este criterio evitamos problemas innecesarios como: gastos computacionales, dificultad de interpretar los resultados o incluso que el árbol no trabaje bien con nuevos datos. Entre los criterios comunes de parada podemos encontrar:

- La pureza del nodo: Si el *split* resultante no satisface la desigualdad 2.4 entonces el algoritmo clasificará al nodo como *terminal*.
- Número de observaciones en el nodo terminal. La división realizada sobre un nodo tienen que dar como resultado un número mínimo de observaciones en cada *nodo hijo*.
- Número de casos mínimos en el *split*. Es el mínimo número de observaciones con el que el modelo intenta calcular el *split*.
- El parámetro de complejidad del árbol que se encarga de evaluar el coste de incluir otra variable al modelo.

El parámetro de complejidad es el más util pues permite crecer árboles de modo que posteriormente es posible podarlo a interés del analista mediante la función *prune()* de R, [7].

2.4.3. Valores perdidos

Cuando la variable usada en el *split* contiene un valor perdido, el árbol toma una variable auxiliar como suplente (*surrogate*) que es, grosso modo, la variable que mejor predice la división establecida en dicho *split*. Se suelen seleccionar varias variables suplentes de forma que cuando llega un caso con un valor faltante en la variable que define el *split*, se usa la primera suplente; si el valor de esta tampoco ha sido observado, se pasa a la siguiente; y si hay valores faltantes en todas las variables suplentes, entonces se decide si descartar el caso o clasificarlo en la clase que mayor probabilidad tiene en ese nodo.

En base a este nuevo concepto podemos definir cómo se calcula la importancia de las variables. Una variable aparece en el árbol varias veces, ya sea como primaria o como suplente. Para obtener la importancia de una variable se suma el valor de la medida de bondad del ajuste de esta variable cada vez que aparece como primaria, más la medida de bondad ajustada por cada vez que aparece como suplente (*surrogate*). Este valor se encuentra escalado hasta 100.

2.5. Particionamiento recursivo basado en modelos

Esta técnica se aplica para encontrar modelos segmentados. La idea básica de un modelo segmentado en este contexto consiste en construir un árbol de clasificación en el que para clasificar a los individuos en un nodo se utiliza un modelo de regresión logística en lugar de utilizar la clase más frecuente. Así, pues, un árbol de clasificación es un caso particular de modelo segmentado donde en cada nodo se ajusta un modelo de regresión logística sin considerar ninguna variable explicativa (solo la constante), y un modelo de regresión logística sería un modelo segmentado donde no hay ninguna variable particionante. Por tanto, se trata de una extensión de las técnicas tratadas anteriormente.

Para construir un modelo segmentado, en primer lugar se eligen las variables que van a formar parte del modelo de regresión logística que se ajustará con los individuos de cada nodo. El resto de las variables serán las candidatas que segmenten o particionen a los individuos. En la referencia [8] explica el algoritmo que emplea estos modelos:

- **Paso 1**: Se ajusta el modelo de regresión logística con los individuos del nodo actual.
- **Paso 2**: Se determina la estabilidad de los parámetros para cada posible variable particionante Z_j, $j = 1, \ldots, l$. Si hay inestabilidad, se elige la variable que más desestabiliza la estimación de los parámetros. En otro caso se para el algoritmo.
- **Paso 3**: Se calcula el punto de segmentación con respecto a la variable que se ha elegido en **Paso 2**.
- **Paso 4**: Se divide el nodo en dos nodos hijos y se regresa al **Paso 1** en cada uno de ellos.

El **Paso 2** del algoritmo está basado en que la estimación en un modelo de regresión logística se ha realizado por máxima verosimilitud, así los valores de los parámetros β verifican

$$\sum_{i=1}^{N} \psi(Y_i, \hat{\theta}) = 0 \tag{2.6}$$

donde ψ es la derivada de $log(y_i log p_i + (1-y_i) log(1-p_i))$ respecto de beta. La idea para encontrar inestabilidad de los parámetros es comprobar si $\hat{\psi}_i$ fluctúa de forma aleatoria alrededor de 0 o si hay una desviación sistemática. Esta desviación alrededor de 0 se mide a través de las sumas parciales de la ecuación 2.6 ordenadas según los valores de la variable particionante.

Una vez seleccionada la variable que más fluctúa, se busca el punto de particionamiento (**Paso 3** del algoritmo) mediante técnicas basadas en el menor error del modelo.

2.6. Validación cruzada

La validación cruzada es una técnica empleada para que los resultados de un análisis estadístico sean lo más independientes posible del conjunto de datos. Esta técnica evita problemas de sobreajuste de los parámetros del modelo a los datos.

La validación cruzada (CV) más común es la denominada k-fold CV, con k=10, que consiste en dividir al azar el conjunto de datos en k partes iguales y crear k modelos, eliminando en cada uno de ellos un conjunto de datos que posteriormente se usará para validar el modelo. Este procedimiento permite estimar el error de predicción mediante el promedio de los errores evaluados en el conjunto no considerado en la construcción del modelo, además de dar una estimación de cada uno de los parámetros.

2.7. Curvas ROC

Las curvas ROC (*Receiver Operating Characteristic curve*) son representaciones gráficas que permiten estudiar la bondad de un modelo matemático predictivo (consultar la referencia [9]). Para entender estas gráficas (véase 2.2) es necesario explicar cómo actúa un clasificador binario cuando predice sobre un conjunto de datos. El modelo será evaluado mediante una matriz de confusión, una tabla bidimensional, donde se enfrentan las predicciones frente a las categorías reales, tabla 2.1.

		Valores reales		
		Verdadero	Falso	Total
Predicciones	Verdadero	Verdaderos positivos	Falsos positivos	P’
	Falso	Falsos negativos	Verdaderos negativos	N’
	Total	P	N	

Tabla 2.1: Matriz de confusión.

El confrontamiento entre los valores reales y predicciones da lugar a las siguientes situaciones:

- **Verdaderos positivos (VP)**: Clasificados correctamente en la clase de interés.
- **Verdaderos negativos (VN)**: Clasificados correctamente en la clase que no es de interés.
- **Falsos postivos (FP)**: Clasificados incorrectamente en la clase de interés.
- **Falsos negativos (FN)** Clasificados incorrectamente en la clase que no es de interés.

Esta matriz de confusión nos ofrece medidas que nos permiten evaluar el comportamiento de un modelo, siendo estas medidas:

- Exactitud o tasa de aciertos (ACC):

$$Exactitud = \frac{VP+VN}{VP+VN+FP+FN}$$

- Tasa de error:

$$Tasa\ de\ error\ = \frac{FP+FN}{VP+VN+FP+FN} = 1 - Exactitud$$

- Ratio o razón de falsos positivos (FPR):

$$FPR = \frac{FP}{N} = \frac{FP}{FP+VN}$$

- Precisión:

$$Precision = \frac{VP}{VP+FP}$$

- Sensibilidad o razón de verdaderos positivos (VPR):

$$VPR = \frac{VP}{P} = \frac{VP}{VP+FN}$$

- Especifidad o razón de verdaderos negativos (SPC):

$$SPC = \frac{VN}{N} = \frac{VN}{FP+VN} = 1 - FPR$$

Estos valores pueden ser visualizados mediante la curva de ROC que enfrenta la *sensibilidad* frente a $1 - Especifidad$ creando el espacio de ROC, ver figura 2.2. El mejor método posible de predicción se situaría en un punto en la esquina superior izquierda, o coordenada (0,1) del espacio ROC y una área bajo la curva igual a la unidad (AUC, *Area under the Curve*), representando un 100 % de sensibilidad (ningún falso negativo) y un 100 % también de especificidad (ningún falso positivo). Por el contrario, una clasificación totalmente aleatoria daría un punto a lo largo de la diagonal, denominada línea de no-discriminación.

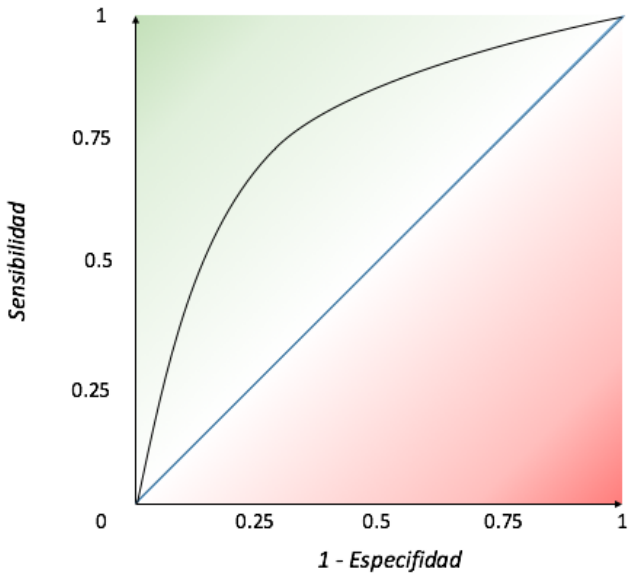

Figura 2.2: Ejemplos de curva de ROC. La curva en la zona verde significa que es un buen clasificador. Por el contrario, la curva de la zona roja indica que es un mal clasificador.

Capítulo 3

Análisis exploratorio de los datos

"Mejor que de nuestro juicio,
debemos fiarnos del cálculo algebraico"

Leonhard Euler

3.1. Introducción

El análisis exploratorio de los datos tiene como objetivo primordial resumir y visualizar los datos de manera que facilite la identificación de patrones que son relevantes para responder a una pregunta de interés. Este análisis se basa en estudiar los gráficos y los estadísticos que permiten entender la distribución de los datos e identificar las importantes características, tales como: concentraciones de valores, forma de la distribución, valores atípicos o outliers...

En esta sección llevaremos a cabo una depuración de los datos lo que nos permitirá continuar con el estudio. A continuación, añadiremos nuevas variables de interés desde un punto de vista clínico, definidas a partir de las anteriores. Realizaremos un análisis descriptivo tanto numérico como gráfico y concluiremos con un estudio de la relación de cada una de las variables con el grupo E4, mediante los test habituales paramétricos y no paramétricos.

3.2. Depuración

Dos cuestiones relevantes son estudiadas en esta sección: los casos duplicados y los datos faltantes. En el primer caso nos percatamos de que la base de datos tiene más de una observación para un sólo individuo, esto se debe a que cuando registraron los datos lo hicieron en dos turnos, en vez de uno: primero se completaron los campos obtenidos por la analítica y después, los relativos CRD (Cuaderno de Recogida de Datos). Por otro lado existen datos faltantes en algunas de las variables, en este caso se ha decido tomar las medidas pertinentes que comentaremos a continuación.

3.2.1. Casos duplicados

Antes de comenzar con el preprocesamiento de los datos hay que comprobar si existen casos duplicados, identificando aquellas líneas que aparecen con un mismo número de secuencia *SEQN*. Como resultado destacamos un total de 52 *SEQN* duplicadas, es decir, tenemos un total de 104 líneas en el conjunto de datos que hay que analizar detenidamente y cuyos casos recogemos en la tabla 3.1.

Tras tomar las medidas pertinentes que se muestran en la tabla 3.1 nos quedamos con un total de 55 casos no aprovechables.

Situación	Caso	Acción / Medida
1	Hay más variables observadas en uno que en otro	Nos quedamos con el que más tiene
2	Faltan todos los valores correspondientes a CRD	Se eliminan los dos
3	En los dos hay el mismo número de variables no observadas pero en uno son de analíticas y en el otro de CRD	Nos quedamos con el que menos faltantes tenga en analíticas
4	Ninguno de los anteriores	Nos quedamos con uno de los dos

Tabla 3.1: Situaciones de los casos duplicados.

3.2.2. Valores perdidos

La variable *ApoE* presenta un total de 44 casos perdidos fácilmente identificables ya que las variables *snp158* y *snp112* (abreviaciones de las secuencias genéticas) son igual a CT. Estos valores perdidos tienen una explicación sencilla y es porque los expertos no lo clasifican en ninguno de los grupos considerados.

En segundo lugar nos centraremos en estudiar la participación de los valores perdidos sobre las variables obteniendo sus porcentajes, tabla 3.2.

ApolipoproteínaA1	ApolipoproteínaB	Lipoproteína.a	TSH	ProteínaCReactiva
0,0026	0,0010	0,2165	0,0003	0,0921
HemoglobinaGlicosilada	**CreatininaEnOrina**	**Microalbumina**	**dtCD**	**PesoTrabajador**
0,0003	0,0803	0,1328	0,0196	0,0211
PerimetroAbdominal	**HoraMuestra**	**PresionSistolica**	**PresionDiastolica**	**PulsacionesPorMinuto**
0,0196	0,0432	0,0229	0,0229	0,0229
ConsumoAlcohol	**ApoE**	**MedicaDiabetes**	**MedicaDislipemia**	**MedicaHipertension**
0,0345	0,0113	0,0247	0,0358	0,0196
MedicaAntiagregante				
0,0201				

Tabla 3.2: Proporciones de valores perdidos en una columna.

La mayor parte de las columnas tienen valores perdidos y en especial las columnas de *Lipoproteína.a* y *Microalbumina*. Generalmente no existe la posibilidad de saber a qué se deben los valores perdidos de un conjunto de datos, pero aquí sí ya que existen columnas comentario asociadas a algunas variables, por ejemplo *ApolipoproteínaA1*, *ApolipoproteínaB*, *ProteínaCReactiva*, *Microalbumina* y *T4libre*.

Estudiemos las variables con mayor número de datos faltantes según los valores de la variable *ApoE* con objeto de determinar si esto afecta a la representatividad de alguno de los genotipos.

Tabla de frecuencias	*ApoE*			
Lipoproteína.a.Comment	**E2**	**E3**	**E4**	**Recuento**
Fuera de rango inferior	33	258	58	349
No informado	51	327	82	460
null	0	3	0	3

Tabla de frecuencias	*ApoE*			
Microalbumina.Comment	**E2**	**E3**	**E4**	**Recuento**
Fuera de rango inferior	54	328	89	471
Fuera de rango superior	0	1	0	1
No informado	2	22	6	30

Porcentaje en filas:	*ApoE*			
Lipoproteína.a.Comment	**E2**	**E3**	**E4**	**Total**
Fuera de rango inferior	9.5	73.9	16.6	100
No informado	11.1	71.1	17.8	100
null	0.0	100.0	0.0	100

Porcentaje en filas:	*ApoE*			
Microalbumina.Comment	**E2**	**E3**	**E4**	**Total**
Fuera de rango inferior	11.5	69.6	18.9	100
Fuera de rango superior	0.0	100.0	0.0	100
No informado	6.7	73.3	20.0	100

Tabla 3.3: Distribución de las columnas comentarios según *ApoE* para *Lipoproteína* y *Microalbumina*.

Aunque la distribución de los datos faltantes es coherente con el porcentaje de individuos esperado en cada uno de los grupos de *ApoE*, se ha decidido tratar de paliar el problema sustituyendo los valores etiquetados como 'Fuera de rango inferior' (son aquéllos no detectados en la analítica por ser más bajos que lo que es capaz de detectar el método) por el mínimo, puesto que esto no distorsionaría los analisis posteriores. De los valores etiquetados como 'No informado' no se tiene información para

su posible imputación y etiquetados como 'Fuera del rango superior' solo hay un caso en la variable *Microalbumina* y otro en la variable *ProteínaCReactiva* que se ha decidido no imputar.

Además se estudian los valores perdidos por procedencia: variables procedentes de analíticas o de CRD. La intención es descartar casos con un alto contenido de valores perdidos, como por ejemplo, ausencia de todos los valores de unos de los grupos de variables, o su mayoría. Son identificados un total de 76 casos que no tienen ningún valor de CRD y por tanto también son retirados del estudio. En resumen, tenemos un conjunto de 3764 observaciones y 37 variables.

Tabla de frecuencias:	*ApoE*			
ProteínaCReactiva.Comment	E2	E3	E4	Recuento
Fuera de rango inferior	27	231	90	348
No informado	0	2	0	2

Porcentaje en filas:	*ApoE*			
ProteínaCReactiva.Comment	E2	E3	E4	Total
Fuera de rango inferior	7.8	66.4	25.9	100.0
No informado	0.0	100.0	0.0	100.0

Tabla 3.4: Distribución de las columnas comentarios según *ApoE* para *ProteínaCReactiva*.

3.3. Obtención de las nuevas variables

Como última preparación de la base de datos incluimos nuevas variables que pueden ser relevantes. En primer lugar creamos variables utilizadas habitualmente en la práctica clínica y que proporcionan en muchos casos más información que las variables tal y como se han recogido originalmente. En segundo lugar crearemos las variables que resultan del producto de la variable dicotómica, que mide si el individuo toma medicación para alguna enfermedad, con una variable numérica, ya que estas variables pueden ser de interés para demostrar la influencia de un medicamento sobre un factor de riesgo.

- **Indice de Masa Corporal (IMC)**: este índice evalúa si el peso que tenemos es adecuado a nuestra altura.

$$ICM = \frac{Peso}{Altura^2}$$

  Esta variable se puede dividir en las siguientes categorías: infrapeso=IMC<18.5, Normal=18.5-24.9 y Obesidad>=25.

- **Indice de Cintura Altura (ica)**: resulta del cociente entre la cintura y la altura, en centímetros. Esta variable sirve para evaluar el riesgo cardiovascular y el estado nutricional de la persona.

$$ICA = \frac{Cintura}{Altura}$$

  Es posible catalogar esta variable de la siguiente forma: Delgado = *ICA*<42, normal = 42-49 y sobrepeso = *ICA*>49.

- **Ratio entre Apolipoproteínas** (*ApoBApoA1*): Esta nueva variable resulta del cociente entre Apolipoproteína B y apolipoproteína A1. Suele ser un mejor indicador que las dos variables por separado.

$$ApoBApoA1 = \frac{ApolipoprotenaB}{ApolipoprotenaA1}$$

- **Ratio Colesterol**:

$$RatioColesterol = \frac{Colesterol}{HDL.Colesterol}$$

A continuación, incluimos las interacciones de las variables con los indicadores que miden si un individuo toma medicación para una determinada enfermedad (diabetes, hipertensión, displemia, formación de trombos y síndrome metabólico).

Por último crearemos las variables categóricas asociadas a cada una de las mediciones analíticas, teniendo en cuenta las agrupaciones mencionadas en el apartado 1.3, para ver si se manifiesta algún comportamiento diferente al observado por la variable numérica. Estas variables aparecen registradas en la base de datos con el sufijo '<NombreVariable>.EF'.

3.4. Análisis descriptivo numérico y gráfico

En primer lugar, realizaremos un estudio descriptivo para observar si hay variables con un comportamiento claramente diferente en cada grupo de *ApoE*. Aunque el objetivo principal es seleccionar las variables relacionadas con el grupo E4, hemos realizado el primer estudio comparativo con los tres grupos de *ApoE* para tener mayor información y plantear si un estudio posterior para E2 podría ser adecuado.

Para este análisis descriptivo se han calculado las medidas habituales de posición, dispersión y forma de las variables cuantitativas, tabla 3.5, y se han representado los diagramas de caja y funciones de densidad estimadas por el método de kernel gaussiano con una anchura de banda dado por la regla de Silverman. La figura 3.1 muestra el gráfico de densidad y el diagrama de caja de la variable *Apolipoproteína B* según los valores de *ApoE*. El estudio de las variables restantes se puede encuentrar en el Anexo A.1.

ApoE	mean	sd	se(mean)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	data:n	data:NA
E2	93.7887	23.3796	1.1869	32	0.2493	0.1291	-0.0235	41	78	94	110	181	388	1
E3	106.5990	25.1298	0.4838	34	0.2357	0.2820	-0.0282	36	89	105	123	215	2698	2
E4	110.9246	24.3735	0.9388	31.55	0.2197	0.2109	0.1461	39	94.45	110	126	211	674	1

Tabla 3.5: Ejemplo de resumen numérico.

De este estudio cabe destacar que parece haber diferencias en las variables *Colesterol*, *HDL.Colesterol*, *GGT*, *Trigliceridos*, *Glucosa*, *ApoA1*, *ApoB*, *Lipo(a)* y *Microalbumina*, presentando el grupo E4 valores mayores que los otros dos en todos los casos salvo *HDL.Colesterol* y *ApoA1* que presenta valores menores y en *Trigliceridos* que toma un valor entre los de los otros dos grupos. La variabilidad comparada con el coeficiente de variación es muy similar en todas ellas. También se ha observado la forma de los datos destacando variables como *BilirrubinaTotal*, *Creatinina*, *GGT*, *Glucosa*, *ALT.GPT*, *Trigliceridos*, *ProteínaCReactiva*, *TSH*, *HemoglobinaGlicosilada* y *Microalbumina* por tener valores elevados de curtosis y asimetría estadística y por lo tanto no guardan parecido la distribución normal. Los gráficos de las demás variables se pueden encontrar en el Anexo A.2.

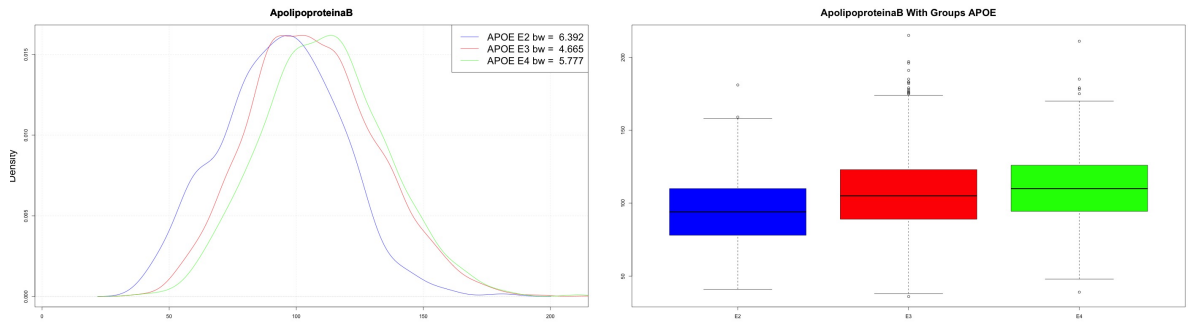

Figura 3.1: Gráfico de densidad y diagrama de caja a la derecha ambos para la variable *ApolipoproteínaB*.

Por último nos queda por estudiar las variables categóricas según el grupo de apolipoproteína utili-

zando los test de chi-cuadrado de Pearson, de este modo sabremos si existe alguna dependencia. En este caso solo la variable *MedicaDisplemia* acepta la dependencia con el grupo *ApoE*.

Como cabría de esperar ninguna variable por sí sola es capaz de distinguir los grupos de *ApoE* pues las medidas obtenidas (media y dispersión) muestran que las distribuciones de cada uno de los grupos de genotipos se solapan entre sí. Por tanto se ha decido que es más relevante identificar las variables presentes en los individuos con *ApoE4*.

3.5. Comparación de grupos

En esta sección se va a estudiar el comportamiento de cada una de las variables cuantitativas en el grupo E4 frente al resto. Para ello se va a utilizar un contraste de hipótesis para confirmar o refutar las diferencias detectadas en el análisis descriptivo.

Se va a utilizar tanto el test paramétrico *t* de comparación de medias de muestras independientes (Tabla 3.6), puesto que se dispone de muestras grandes, aunque se va a aplicar también el correspondiente test no paramétrico de Mann-Whitney-Wilcoson (Tabla 3.7).

Vemos que existen diferencias significativas en las variables *Colesterol*, *HDL.Colesterol*, *apolipoproteína A1*, *apolipoproteína B*, *Lipoproteína.a*, *ProteínaCReactiva*, *ApoBApoA1* y *RatioColesterol*, que son algunas de las variables para las que se han visto diferencias en el estudio descriptivo. También se han hecho test de chi-cuadrado para las variables categóricas y en todos los casos no hay evidencia para rechazar la independencia con el grupo de *ApoE4*. Las tablas de doble entrada y los resultados del test se encuentran en el Anexo A.2.2.

Variable	pvalues		Variable	pvalues	
BilirrubinaTotal	0.8103		TSH	0.1053	
Calcio	0.1927		HemoglobinaGlicosilada	0.7150	
Colesterol	0.0181	*	CreatininaEnOrina	0.1938	
HDL.Colesterol	0.0070	**	Microalbumina	0.3592	
Creatinina	0.4773		PesoTrabajador	0.3931	
GGT	0.2092		PerimetroAbdominal	0.4622	
Glucosa	0.4023		PresionSistolica	0.1492	
AST.GOT	0.3006		PresionDiastolica	0.5768	
ALT.GPT	0.6376		PulsacionesPorMinuto	0.7862	
Trigliceridos	0.1896		ConsumoAlcohol	0.1815	
AcidoUrico	0.8064		ApoBApoA1	0	***
ApolipoproteínaA1	0.0017	**	RatioColesterol	0.0460	*
ApolipoproteínaB	0.0001	**	Altura	0.4519	
Lipoproteína.a	0.0032	**	Edad	0.6009	
ProteínaCReactiva	0.0009	**	ICM	0.6080	
T4libre	0.8939		ica	0.6562	

*** p-valor <0.0001 ** p-valor <0.01 * p-valor<0.05

Tabla 3.6: Test de hipótesis paramétrico *t* de student.

Variable	pvalues		Variable	pvalues	
BilirrubinaTotal	0.6456		TSH	0.8575	
Calcio	0.1733		HemoglobinaGlicosilada	0.8232	
Colesterol	0.0142	*	CreatininaEnOrina	0.2347	
HDL.Colesterol	0.0014	**	Microalbumina	0.8218	
Creatinina	0.9608		PesoTrabajador	0.7274	
GGT	0.6593		PerimetroAbdominal	0.5394	
Glucosa	0.4817		PresionSistol ICA	0.2292	
AST.GOT	0.3222		PresionDiastol HDL.Colesterol	0.6341	
ALT.GPT	0.7092		PulsacionesPorMinuto	0.5812	
Trigliceridos	0.7021		ConsumoAlcohol	0.2208	
AcidoUrico	0.7862		ApoBApoA1	0	***
ApolipoproteínaA1	0.0007	**	RatioColesterol	0.0001	**
ApolipoproteínaB	0	***	Altura	0.5015	
Lipoproteína.a	0.0073	**	Edad	0.5185	
ProteínaCReactiva	0.0001	**	ICM	0.8410	
T4libre	0.9177		*ICA*	0.7162	

*** p-valor < 0.0001 ** p-valor <0.01 * p-valor<0.05

Tabla 3.7: Test de hipótesis no paramétrico Mann-Whitney-Wilcoxon.

3.6. Correlación entre las variables

Una matriz de correlación ayudará a determinar la relaciones entre parejas o grupos de variables del conjunto de datos. La figura 3.2 muestra la existencia de correlaciones entre las variables como: *Colesterol*, *Apoliproproteína B* y *ApoBApoA1*; *ApolipoproteínaA1* y *HDL.Colesterol*; *Hemoglobina-Glicada* y *Glucosa*; *ICM*, *HDL.Colesterol*, *PesoTrabajador* y *PerímetroAbdominal*; *PresiónDiastolica* y *PresionSistolica* y finalmente *AST.GOT* y *ALT.GPT*.

A su vez existen variables con una correlación negativa como por ejemplo *Triglicéridos* y *HDL.Colesterol*, esta correlación es clara: cuando ingerimos mayor cantidad de *Triglicéridos*, grasas, tenemos menos HDL, colesterol 'el bueno' que se encarga de transportar el colesterol desde los tejidos hasta hígado para proceder con su eliminación.

Hay una relación obvia entre las variables de *Apolipoproteína A1* y *HDL.Colesterol* pues, la apolipoproteína A-I (Apo A-I) es el principal componente proteico de las lipoproteínas de alta densidad, el colesterol el bueno (*HDL.Colesterol*), por tanto si una variable aumenta consecuentemente se incrementará la correlacionada. A su vez existe una relación positiva entre *Apoliproproteína B* y *Colesterol*, se debe a apoliproproteína B se encarga de mover el colesterol por la sangre. Del mismo modo encontramos la *HemoglobinaGlicosilada* y la *Glucosa*, ya que la *Glucosa* se une a la hemoglobina para formar la hemoglobina glicosilada. La hemoglobina glicosilada refleja la exposición de las células al azúcar. Por tanto si hay mayor *Glucosa* en sangre tendremos valores elevados de hemoglobina glicosilada.

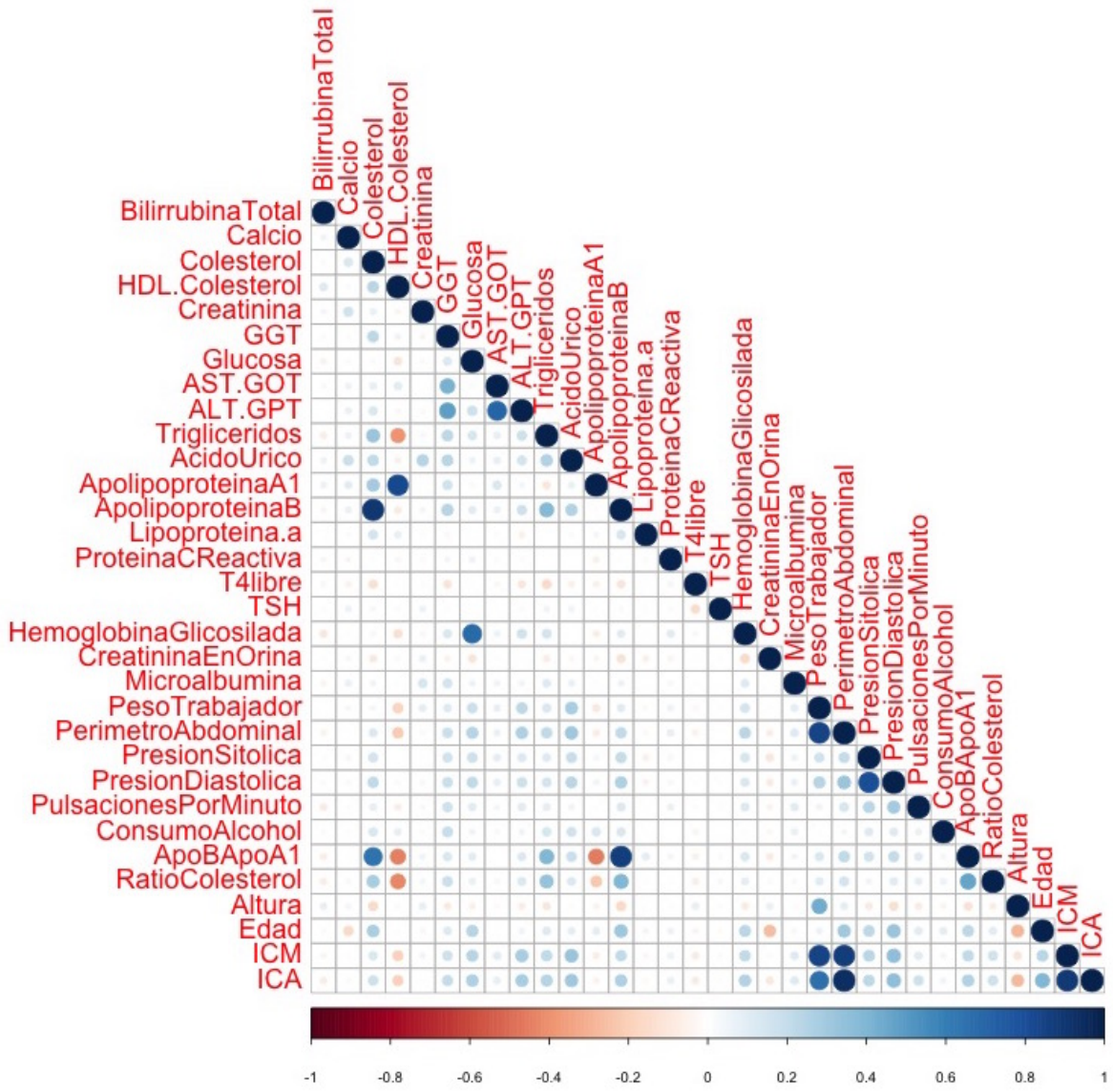

Figura 3.2: Matriz de correlación para las variables numéricas.

Capítulo 4

Modelos

"Si buscas resultados distintos
no hagas siempre lo mismo"

Albert Einstein

4.1. Introducción

A continuación se exponen los modelos que han sido elaborados mediante las distintas técnicas explicadas en el capítulo 2. Se comenzará exponiendo los resultados de las técnicas más comunes para posteriormente pasar a la regresión logística penalizada y al particionamiento recursivo basado en modelos.

Este capítulo comienza exponiendo los resultados de la regresión logística combinada con la validación cruzada, que se ha realizado a mano para poder explicar dicho proceso con un ejemplo. Seguidamente son introducidos los resultados de los árboles de clasificación que primeramente han sido crecidos con un parámetro de complejidad nulo ($c_p = 0$) para permitir observar cuáles variables son más relevantes y cuáles son prescindibles. Después en los resultados de la regresión logística penalizada se podrá observar si existe algún efecto de interacción entre las variables categóricas creadas o las variables originales sobre las originales. A continuación se podrá observar los resultados del particionamiento recursivo basado en modelos de regresión logística. Finalizando este capítulo con una conclusión sobre los resultados obtenidos.

4.2. Regresión logística

Recordemos que datos vamos a utilizar para crear estos modelos: variables originales, variables resultantes de utilizar la categorización mencionada en la sección 1.3[1] y las interacciones entre las variables dicotómicas con las numéricas.

Utilizaremos la función *stepwise* de R lo que nos permitirá realizar una regresión paso a paso. Una vez elegido el procedimiento pasamos a realizar diversos modelos mediante validación cruzada. Con una fragmentación de los datos en diez grupos se ha creado cada uno de los modelos de CV, en donde se emplea el 90 % de los datos para entrenamiento y el 10 % restante para validación.

(Intercept)	ApoBApoA1	ColesterolEF[T.Normal]	ColesterolEF[T.Superior]	Lipoproteína.a	proteínaCReactiva
10	10	10	10	10	10
ConsumoAlcohol.MDis	**Microalbumina.metSin**	**PresionSitolica**	**proteínaCReactivaEF[T.Riesgo]**	**CreatininaEnOrina**	**BilirrubinaTotalEF[T.Superior]**
9	8	7	7	6	4

Tabla 4.1: Frecuencia de las variables utilizadas en los modelos.

[1] Recordemos que estas variables aparecen etiquetadas de la forma '<NombreVariable>.EF'

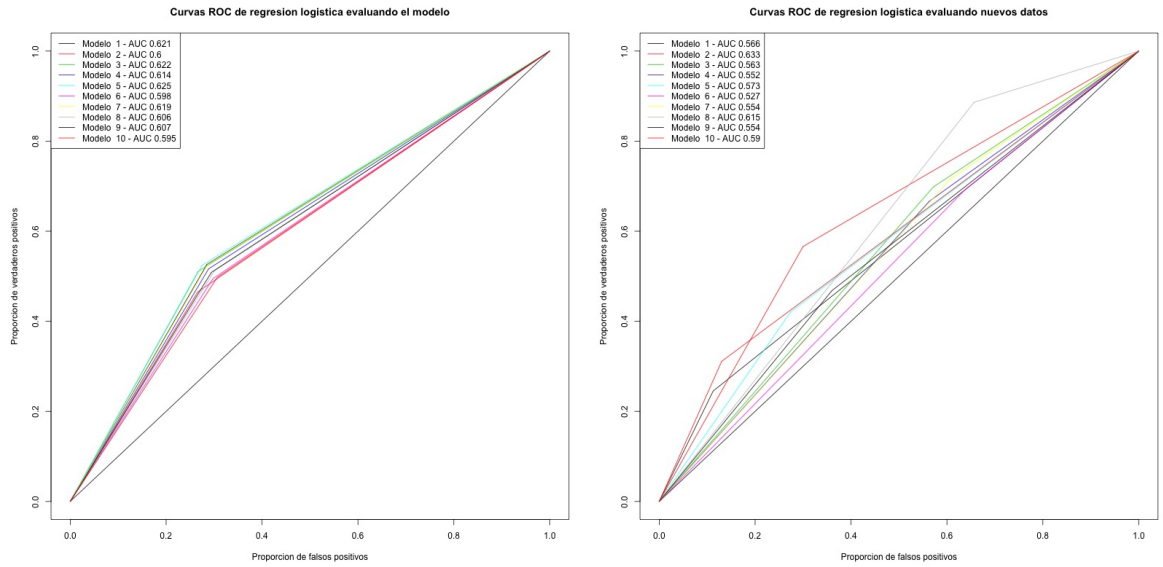

Figura 4.1: Curvas ROC con datos de entrenamiento, izquierda, y validación, derecha.

En un primer modelo nos quedamos con las variables que aparecen más de 7 veces en la tabla 4.1. En la tabla 4.2 podemos ver las variables *ColesterolEF[T.Alto]* y *ProteínaCReactivaEF[T.Alto]* que han sido creadas por R. Estas variables corresponden a la categoría *Alto* de las variables *ColesterolEF* y *ProteínaCReactivaEF*.

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-2.808	0.461	-6.087	0.000	***
ApoBApoA1	1.273	0.269	4.739	0.000	***
ColesterolEF[T.Alto]	-0.052	0.117	-0.446	0.655	
Lipoproteína.a	0.004	0.002	2.285	0.022	*
proteínaCReactiva	-1.085	0.338	-3.210	0.001	**
ConsumoAlcohol.MDis	0.003	0.002	2.150	0.032	*
PresionSitolica	0.004	0.003	1.075	0.282	
Microalbumina.metSin	0.002	0.001	1.402	0.161	
proteínaCReactivaEF[T.Alto]	0.057	0.232	0.247	0.805	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tabla 4.2: Primer modelo con variables de mayor frecuencia.

Hasta este punto tenemos una base del modelo, ahora nos centraremos en mejorarlo. En la búsqueda de este nuevo modelo se han ido retirando variables una por una (según el p-valor más elevado). excluyéndolas en el siguiente orden *ProteínaCReactivaEFT.Alto]*, *ColesterolEF[T.Alto]*, *PresionSitolica*, *Microalbumina.metSin* y *ConsumoAlcohol.MDis*. Los modelos auxiliares se pueden encontrar en el Anexo B.1. Del modelo final, ver tabla 4.3 podemos destacar:

- *ProteínaCReactiva*: Esta variable aparece con término negativo. Es llamativo pues debería salir positivo ya que a mayor cantidad implica poder desarrollar enfermedades cardiovasculares y por ende, un aumento en la probabilidad de pertenecer al grupo E4.

- *Lipoproteína.a*: tienen coeficiente positivo, luego los valores altos afectan en la probabilidad de pertenecer al grupo E4. Según [5] nos dice: 'Los valores de lp (a) superiores a lo normal están asociados con un alto riesgo de aterosclerosis, accidente cerebrovascular y ataque cardíaco.' Por tanto consideramos normal que este factor influya en la probabilidad del grupo E4.

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-2.344	0.188	-12.447	<2E-16	***
ApoBApoA1	1.278	0.229	5.578	0.000	***
Lipoproteína.a	0.003	0.002	2.201	0.028	*
proteínaCReactiva	-0.955	0.211	-4.525	0.000	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tabla 4.3: Modelo elegido.

- *ApoBApoA1*: tiene coeficiente positivo, los valores elevados señalan la pertenencia al grupo E4. Existen dos posibilidades (a excepción de las triviales) para que este valor sea alto:

  1. El valor elevado de ApoB mientras el ApoA1 se mantiene constante. Lo que significa que hay altos niveles de lípidos en el cuerpo.
  2. El pequeño valor ApoA1 mientras ApoB es constante (y en un valor normal). Este valor indica que hay problemas en la activación de la lecitina colesterol aciltransferasa que cataliza la esterificación del colesterol.

  Por tanto en estos dos casos hacen que aumente el valor de esta variable. Y consecuentemente este modelo nos dice que los valores altos indican que afecta positivamente a la probabilidad de pertenecer al grupo E4.

Por consiguiente este es el modelo de regresión logística resultante con una curva ROC, ver figura 4.2.

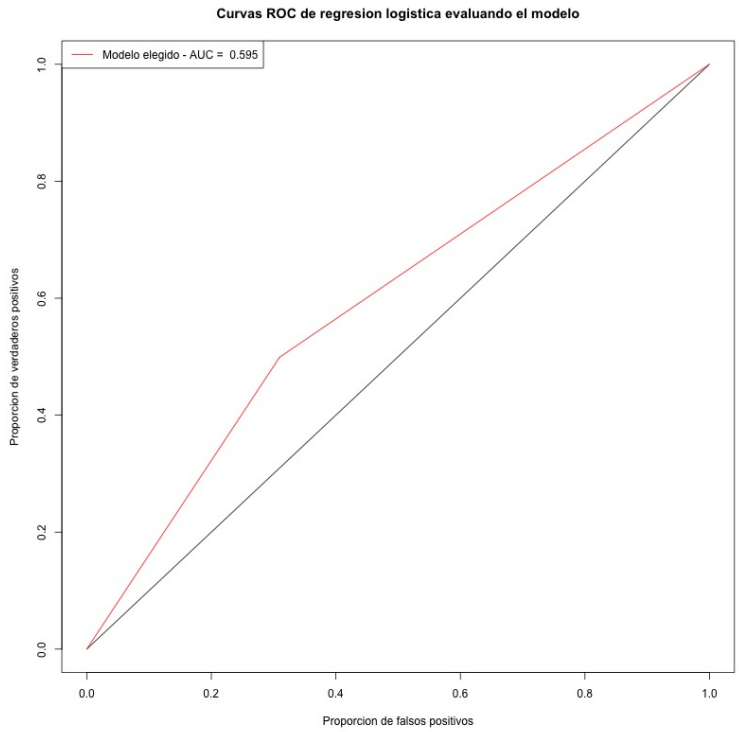

Figura 4.2: Curva ROC del modelo de regresión logística elegido.

4.3. Árboles de clasificación

En primer lugar se han construido los modelos con todas las variables, a excepción de las categorizadas con el sufijo EF, utilizando el criterio de información y $c_p = 0$ con objeto de forzar a la técnica a hacer todas las divisiones posibles y determinar qué variables son las que más veces aparecen formando parte de una división del árbol (*split*) o como suplentes (*surrogate*). Los resultados se muestran en la tabla 4.4 bajo el epígrafe de modelo 1. A la vista de que las variables más importantes están correladas (como se ha explicado en el apartado 3.6), se ha vuelto a construir el modelo en las mismas condiciones que el anterior, pero eliminado aquellas variables que presentaban un coeficiente de correlación por encima de 0.6. Los resultados se muestran en la tabla 4.4 bajo el epígrafe de modelo 2. Finalmente, se ha procedido a podar el árbol (modelo 3 en la tabla 4.4). Este mismo proceso se ha repetido utilizando el criterio de Gini y los resultados se corresponden con los modelos 4, 5 y 6 de la tabla 4.4.

Los modelos 3 y 6 son los árboles ya podados con un mínimo en el error de validación cruzada que corresponde a su vez con un mínimo en el *split*, ver la figura 4.3. En ambos árboles tenemos el primer nodo terminal hacia la izquierda separando la mayor parte de los casos (59 % y 65 %, respectivamente) sin la apolipoproteína E4 con errores de clasificación de $0,14$ y $0,15$. El segundo árbol podemos ver que la clasificación del grupo E4 de los nodos terminales tiene valores cercanos a la unidad siendo la unidad uno de ellos.

En el primer árbol tenemos con mayor importancia la variable *ProteínaCReactiva* con valor del 24 % seguida de *apolipoproteína B*,*Triglicéridos*, *RatioColesterol* y *GGT*. En cambio, en el segundo árbol tenemos *ApoBApoA1* con un valor de 20 % seguido de *ProteínaCReactiva*, *RatioColesterol*, *Triglicéridos*, *Glucosa* y *Creatinina*. Esto nos dice que estás son las variables que más salen como primarias o sustitutas en cada uno de los *splits* realizados. Ambos árboles tienen en común las variables *ProteínaCReactiva*, *Glucosa*, *Triglicéridos* y *RatioColesterol*, participando esta última como sustituta. Los resultados se pueden ver en el Anexo B.2.

Modelo	Criterio	Variables
Modelo 1	Criterio información $c_p = 0$	IMC PerimetroAbdominal ICA RatioColesterol Triglicéridos apolipoproteína B ApoBApoA1
Modelo 2	Criterio información Sin variables correladas	RatioColesterol apolipoproteína B Triglicéridos HDL.Colesterol Glucosa GGT IMC PresionSitolica Microalbumina
Modelo 3	Criterio información Sin variables correladas min split	proteínaCReactiva apolipoproteína B Triglicéridos RatioColesterol GGT HDL.Colesterol Edad T4libre IMC
Modelo 4	Criterio Gini $c_p = 0$	PesoTrabajador RatioColesterol PerimetroAbdominal IMC ApoBApoA1 ALT.GPT Triglicéridos
Modelo 5	Criterio Gini Sin variables correladas	ApoBApoA1 RatioColesterol Triglicéridos ALT.GPT Microalbumina CreatininaEnOrina proteínaCReactiva
Modelo 6	Criterio Gini Sin variables correladas min split	ApoBApoA1 proteínaCReactiva RatioColesterol Triglicéridos Glucosa Creatinina ALT.GPT

Tabla 4.4: Tabla de variables importantes obtenidas en los modelos.

La primera partición realizada en el árbol de la izquierda de la figura 4.3 es con la variable proteínaCReactiva catalogando individuos con valores superiores a 0.11 sin apolipoproteína E4, este valor sale contrario al que nos dice la teoría. De la segunda rama destacaremos el nodo terminal situado a la derecha del árbol, el camino que recorre pasa por valores altos de *apolipoproteína B*, *Triglicéridos* y *T4libre* que, como ya habíamos visto en la sección 1.3, valores elevados implica desarrollo de enfermedades y por tanto pertenecer al grupo E4. Como segundo nodo terminal que clasifica los individuos de E4 tenemos aquellas observaciones con valores elevados de *apolipoproteína B*, bajos en *Triglicéridos*, altos en *Glucosa* y bajos en *HDL.Colesterol*.

En el segundo árbol, la primera partición que realiza hacia la izquierda es con la variable *ApoBApoA1*, clasificando los individuos sin pertenecer al grupo E4. De la segunda rama destacaremos el nodo terminal con una clasificación del 100 % cuyo camino recorre niveles de *ProteínaCReactiva* y *ApoBA-*

poA1 bajos. Nuevamente, *ProteínaCReactiva* sale contrario al esperado pues valores altos implican que hay una inflamación. A partir del *split ProteínaCReactiva* tenemos un nodo terminal con valores elevados de *ApoBApoA1* y *Glucosa* y otro nodo con valores altos de *ApoBApoA1*, bajos de *Glucosa*, y altos en *Triglicéridos* y *Creatinina*. De este árbol también hay que destacar que hay dos *split* con la variable *ApoBApoA1* bajo el mismo valor. Si tenemos en cuenta la salida que nos proporciona R y que se puede encontrar en el anexo B.2, podemos observar que las divisiones las hacen bajo los valores de 0.8168 y 0.8220 respectivamente, por tanto los valores que ofrece las figura 4.3 se encuentran redondeados al valor de 0.82.

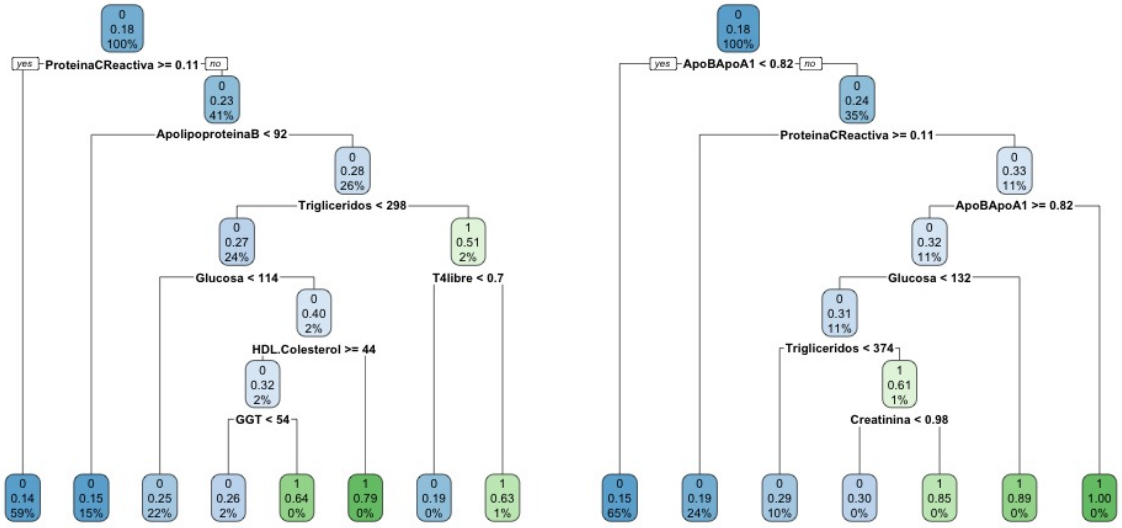

Figura 4.3: Árboles de clasificación crecidos bajo el criterio de Información y Gini, respectivamente y podados.

Las curvas ROC de estos modelos se pueden ver en la figura 4.4.

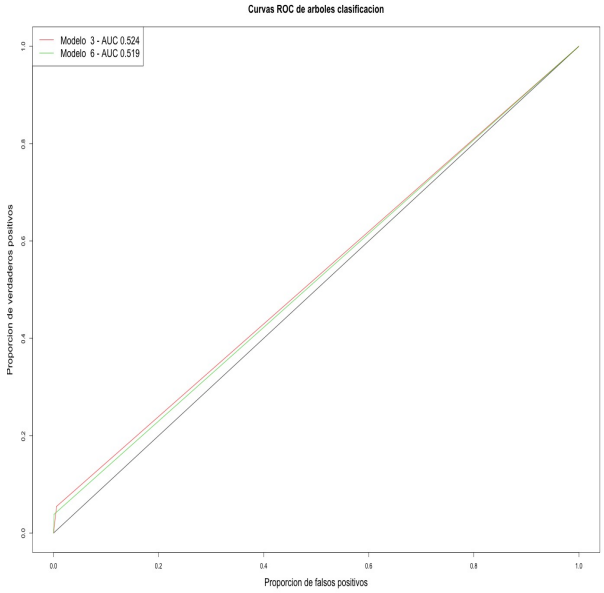

Figura 4.4: Curvas ROC de los modelos de árboles de clasificación.

4.4. Regresión logística penalizada

En la regresión penalizada utilizamos la validación cruzada y distintos valores de α, este parámetro que marca un compromiso entre la regresión ridge ($\alpha = 0$) y la regresión lasso ($\alpha = 1$). La intención es observar el papel que juega la penalización de los coeficientes de beta y cómo influye en el resultado de los modelos. Se han realizado pruebas con todas las variables (tabla 4.6), sin las variables etiquetadas como EF (tabla 4.7) y con las variables originales (tabla 4.8) para ver si influyen de algún modo las variables categóricas EF o las variables de interacción sobre las originales.

Las variables que aparecen en los modelos se encuentran tabla 4.5. Destacaremos *ApoBApoA1*, *ProteínaCReactiva*, *Lipoproteína.a* y *RatioColesterol* por aparecer en todos los modelos y tener coeficientes positivos (a excepción de *ProteínaCReactiva*, ver Anexo B.3) lo que nos indica que valores altos favorecen en la probabilidad de pertenecer al grupo E4. A la vista de las variables más comunes a estos modelos podemos decir que no hay ninguna interacción entre las variables originales y las creadas.

(Intercept)	ApoBApoA1	Lipoproteína.a	proteínaCReactiva	RatioColesterol	apolipoproteína B
15	15	15	15	15	10
ConsumoAlcohol.MDis	**Glucosa.MDis**	**Lipoproteína.a.MAnt**	**Microalbumina.metSin**	**apolipoproteína A1**	**GlucosaEF[T.Alto]**
10	10	10	7	6	5
MedicaDislipemia	**proteínaCReactivaEF[T.Alto]**	**TriglicéridosEF[T.Normal]**	**TSH**	**Lipoproteína.aEF[T.Alto]**	**PresionSitolicaEF[T.Alto]**
5	5	5	4	2	2
BilirrubinaTotalEF[T.Alto]					
1					

Tabla 4.5: Recuento de variables que han aparecen en los modelos de regresión penalizada.

Modelo	Variables
Lasso	(Intercept) Lipoproteína.a proteínaCReactiva ApoBApoA1 RatioColesterol Glucosa.MDis Lipoproteína.a.MAnt ConsumoAlcohol.MDis GlucosaEF[T.Alto] proteínaCReactivaEF[T.Alto] TriglicéridosEF[T.Normal]
CV.net.08	(Intercept) Lipoproteína.a proteínaCReactiva ApoBApoA1 RatioColesterol Glucosa.MDis Lipoproteína.a.MAnt ConsumoAlcohol.MDis GlucosaEF[T.Alto] proteínaCReactivaEF[T.Alto] TriglicéridosEF[T.Normal]
CV.net.06	(Intercept) apolipoproteína B Lipoproteína.a proteínaCReactiva ApoBApoA1 RatioColesterol Glucosa.MDis Lipoproteína.a.MAnt ConsumoAlcohol.MDis GlucosaEF[T.Alto] proteínaCReactivaEF[T.Alto] TriglicéridosEF[T.Normal]
CV.net.04	(Intercept) apolipoproteína A1 apolipoproteína B Lipoproteína.a proteínaCReactiva ApoBApoA1 RatioColesterol Glucosa.MDis Lipoproteína.a.MAnt Microalbumina.metSin ConsumoAlcohol.MDis GlucosaEF[T.Alto] Lipoproteína.aEF[T.Alto] proteínaCReactivaEF[T.Alto] PresionSitolicaEF[T.Alto] TriglicéridosEF[T.Normal]
CV.net.02	(Intercept) apolipoproteína A1 apolipoproteína B Lipoproteína.a proteínaCReactiva TSH ApoBApoA1 RatioColesterol Glucosa.MDis Lipoproteína.a.MAnt Microalbumina.metSin ConsumoAlcohol.MDis BilirrubinaTotalEF[T.Alto] GlucosaEF[T.Alto] Lipoproteína.aEF[T.Alto] proteínaCReactivaEF[T.Alto] PresionSitolicaEF[T.Alto] TriglicéridosEF[T.Normal]

Tabla 4.6: Variables resultantes de los modelos de regresión penalizada con todas las variables.

Modelo	Variables
Lasso	(Intercept) Lipoproteína.a proteínaCReactiva ApoBApoA1 RatioColesterol Glucosa.MDis Lipoproteína.a.MAnt Microalbumina.metSin ConsumoAlcohol.MDis
CV.net.08	(Intercept) Lipoproteína.a proteínaCReactiva ApoBApoA1 RatioColesterol Glucosa.MDis Lipoproteína.a.MAnt Microalbumina.metSin ConsumoAlcohol.MDis
CV.net.06	(Intercept) apolipoproteína B Lipoproteína.a proteínaCReactiva ApoBApoA1 RatioColesterol Glucosa.MDis Lipoproteína.a.MAnt Microalbumina.metSin ConsumoAlcohol.MDis
CV.net.04	(Intercept) apolipoproteína A1 apolipoproteína B Lipoproteína.a proteínaCReactiva ApoBApoA1 RatioColesterol Glucosa.MDis Lipoproteína.a.MAnt Microalbumina.metSin ConsumoAlcohol.MDis
CV.net.02	(Intercept) apolipoproteína A1 apolipoproteína B Lipoproteína.a proteínaCReactiva TSH ApoBApoA1 RatioColesterol Glucosa.MDis Lipoproteína.a.MAnt Microalbumina.metSin ConsumoAlcohol.MDis

Tabla 4.7: Variables resultantes de los modelos de regresión penalizada sin las variables de EF.

Modelo	Variables
Lasso	(Intercept) Lipoproteína.a proteínaCReactiva MedicaDislipemia ApoBApoA1 RatioColesterol
CV.net.08	(Intercept) apolipoproteína B Lipoproteína.a proteínaCReactiva MedicaDislipemia ApoBApoA1 RatioColesterol
CV.net.06	(Intercept) apolipoproteína B Lipoproteína.a proteínaCReactiva MedicaDislipemia ApoBApoA1 RatioColesterol
CV.net.04	(Intercept) apolipoproteína A1 apolipoproteína B Lipoproteína.a proteínaCReactiva TSH MedicaDislipemia ApoBApoA1 RatioColesterol
CV.net.02	(Intercept) apolipoproteína A1 apolipoproteína B Lipoproteína.a proteínaCReactiva TSH MedicaDislipemia ApoBApoA1 RatioColesterol

Tabla 4.8: Variables resultantes de los modelos de regresión penalizada con las variables originales.

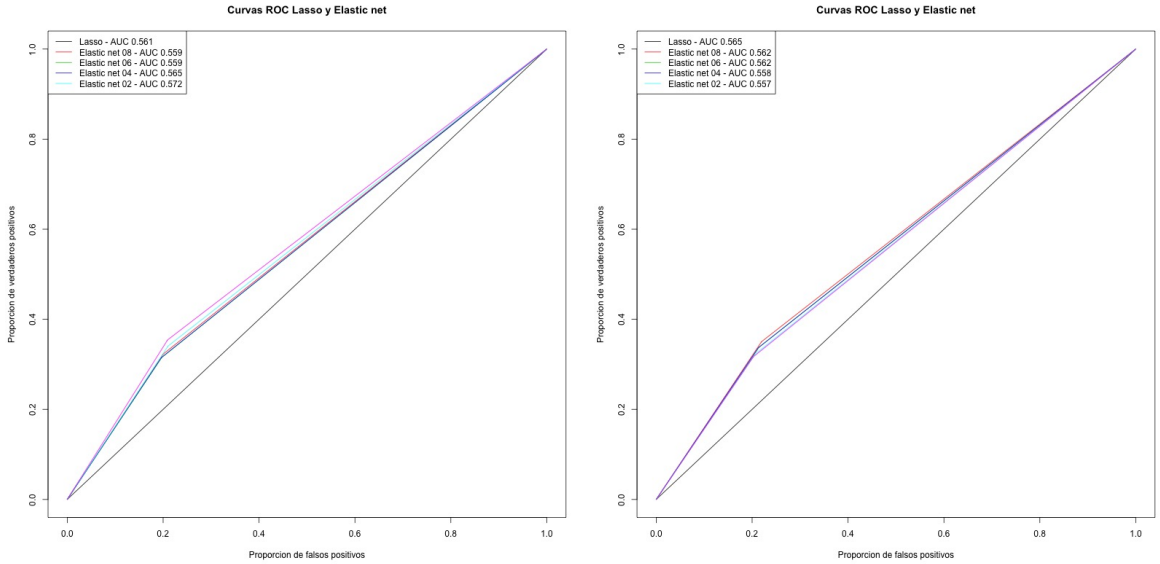

Figura 4.5: Curvas ROC de los modelos de regresión penalizada con todas las variables y sin las recodificadas respectivamente.

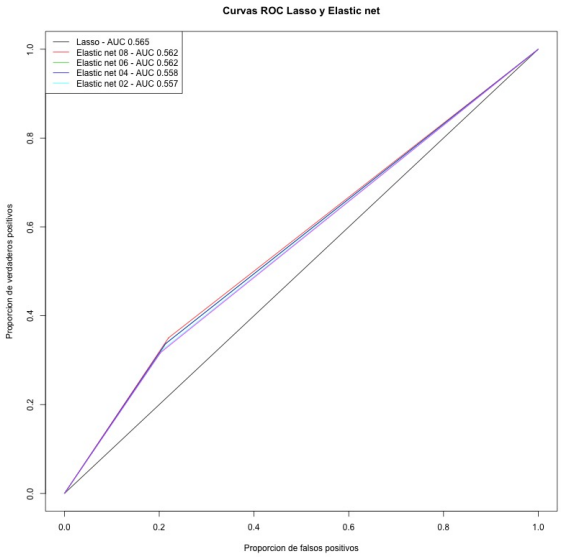

Figura 4.6: Curvas ROC de los modelos de regresión penalizada con las variables originales.

4.5. Particionamiento recursivo basado en modelos

Por último se han realizado modelos de particionamiento recursivo utilizando la regresión logística. En primer lugar se han considerado las variables que han salido en la regresión logística (*ApoBApoA1*, *Lipoproteína.a* y *ProteínaCReactiva*) como regresoras y las restantes como particionamiento. Como el modelo resultante era el mismo que en la regresión logística (solo había un único nodo, ver anexo B.4) se ha decidido considerar como variables de particionamiento a *apolipoproteína B*, *GGT*, *apolipoproteína A1*, *PresionSitolica*, *IMC* y *HemoglobinaGlicosilada* debido a las correlaciones entre los grupos de variables y porque cada una aporta términos médicos de distinta información. Los modelos se muestran en la figura 4.7 y en la tabla 4.9.

Los modelos que aparecen en la figura 4.7 han sido estudiados individualmente. De la primera rama distinguimos individuos con valores de *apolipoproteína B* inferiores a 121 donde se ha quitado la variable *RatioColesterol* ya que tenía un coeficiente negativo, opuesto al que debería salir por lo explicado en la teoría, y por tanto nos quedamos con las variables *ApoBApoA1*, *Lipoproteína.a* y *ProteínaCReactiva*. De la segunda rama no entra ninguna variable ya que ninguna tiene suficiente poder discriminatorio como para marcar diferencias entre los grupos de E4. Los modelos intermedios se pueden encontrar en el Anexo B.4.

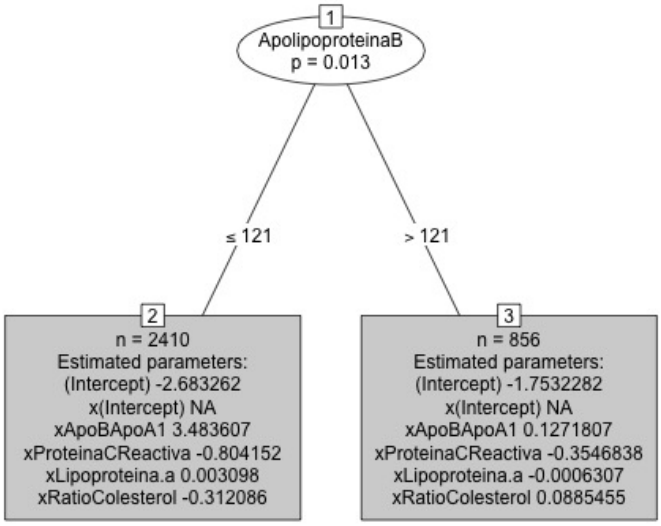

Figura 4.7: Resultados de los modelos recursivos utilizando regresión logística.

Modelo nodo 2

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-2.956	0.268	-11.046	<2E-16	***
ApoBApoA1	2.116	0.371	5.701	0.000	***
Lipoproteína.a	0.004	0.002	2.036	0.042	*
proteínaCReactiva	-0.832	0.222	-3.753	0.000	***

Tabla 4.9: Modelo refinados de 4.7.

4.6. Confrontación de modelos

A lo largo de este trabajo se han explorado diferentes modelos de clasificación, ofreciendo cada técnica resultados diferentes. En la tabla 4.6 se pueden ver un resumen de las variables que han aparecido en los modelos. Destacaremos las variables *ApoBApoA1*, *ProteínaCReactiva* y *Lipoproteína.a* por ser las comunes y por tanto tienen un papel en la determinación del grupo de apolipoproteínas E4. De las variables creadas podemos destacar que no influyen sobre las originales según lo observado en la regresión logística penalizada, ver sección 4.4.

Las variables *ApoBApoA1* y *Lipoproteína.a* presentan una contribución positiva en los modelos de regresión logística, regresión logística penalizada y particionamiento recursivo lo que indica que valores elevados influyen en la determinación del grupo E4. A su vez este resultado concuerda con lo expuesto en la teoría de la sección 1.3, para la variable de *ApoBApoA1* los valores elevados pueden indicar desarrollo de enfermedades como arteriosclerosis, enfermedades circulatorias,... y para *Lipoproteína.a* los valores elevados pueden ser un factor de riesgo de cardiopatía. Por otro lado, la variable *ProteínaCReactiva* presenta una contribución negativa en todos los modelos y para el caso de los árboles un *split* contrario al esperado, pues valores elevados indican que hay una inflamación y por tanto es posible desarrollar enfermedades cardiovasculares. Es un resultado que ya ha aparecido en estudios previos, ver referencia [10], y que por tanto se ha dejado presente en los modelos.

Por último, hay que hacer notar que para valores altos (> 121) del *split* realizado con la variable *ApolipoproteínaB* en los modelos de particionamiento recursivo podemos ver que ninguna de las variables comunes es influyente sobre la determinación del grupo de apolipoproteínas E4.

Modelo	Description	Variables resultantes
Modelo 1 R. Logística	Modelo logístico elegido mediante validación cruzada.	ApoBApoA1, proteínaCReativa, Lipoproteína.a
Modelo 3 Árbol	Criterio Information, min split, sin variables correladas	proteínaCReactiva apolipoproteína B Triglicéridos RatioColesterol GGT HDL.Colesterol Edad T4libre IMC
Modelo 6 Árbol	Criterio Gini, min split, sin variables correladas	ApoBApoA1 proteínaCReactiva RatioColesterol Triglicéridos Glucosa Creatinina ALT.GPT
Regresión penalizada	Modelo logístico penalizado elegido mediante validación cruzada.	ApoBApoA1 proteínaCReactiva Lipoproteína.a RatioColesterol
Partición recursiva	Realizada con la variable apolipoproteína B	ApoBApoA1, proteínaCReactiva, Lipoproteína.a, con apolipoproteína B < 121

Tabla 4.10: Tabla resumen de las variables obtenidas en los distintos métodos.

Bibliografía

[1] Tejedor, M. T., García-Sobreviela, M. P. , Ledesma, M. and Arbonés-Mainar J. M.(2014). The apolipoprotein E polymorphism RS7412 associates with body fatness independently of plasma lipids in middle aged men, *PLOS ONE*, 9 (9),1-5.

[2] Torres-Pérez, E., Ledesma, M., García-Sobreviela, M. P., León-Latre, M. and Arbonés-Mainar, J. M (2015). Apolipoprotein E4 association with metabolic syndrome depends on body fatness, *Atherosclerosis*, 245, 35-42

[3] R, RStudio (2018) Fecha de consulta: mayo, 2017 desde https://www.rstudio.com.

[4] *Wikipedia, la enciclopedia libre.* (2018) Fecha de consulta: noviembre, 2017 desde https://es.wikipedia.org.

- Apolipoproteína E. (2017). *Wikipedia, la enciclopedia libre.* Fecha de consulta: noviembre, 2017 desde https://en.wikipedia.org/wiki/Apolipoprotein_E
- Dislipidemia. (2017). *Wikipedia, la enciclopedia libre.* Fecha de consulta: noviembre, 2017 desde https://es.wikipedia.org/wiki/Dislipidemia
- Antiagregante plaquetario. (2017). *Wikipedia, la enciclopedia libre.* Fecha de consulta: noviembre, 2017 desde https://es.wikipedia.org/wiki/Antiagregante_plaquetario

[5] *U.S. National Library of Medicine.* (2017) Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov.

- Bilirrubina. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003479.html
- Calcio. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003477.html
- Colesterol. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/patientinstructions/000386.html
- Creatinina. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003475.html
- CGT. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003458.html
- Glucemia. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003482.html
- AST. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003472.html
- ALT. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003473.html
- Trigliceridos. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003493.html

- Ácido úrico. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003476.html
- Lipoproteína A. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/007262.html
- proteína C Reactiva. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003356.html
- T4. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003517.html
- TSH. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003684.html
- Hemogoblina. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003645.html
- Microalbuminuria. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/ency/article/003591.html
- Presión arterial. (2017) *U.S. National Library of Medicine.* Fecha de consulta: noviembre, 2017 desde https://medlineplus.gov/spanish/highbloodpressure.html

[6] MayoClinic (1998-2017). Fecha de consulta: noviembre, 2017 desde https://www.mayoclinic.org/

- MayoClinic (1998-2017). Fecha de consulta: noviembre, 2017 desde https://www.mayoclinic.org/es-es/diseases-conditions/liver-problems/symptoms-causes/syc-20374502

[7] Terry M. Therneau and Elizabeth J. Atkinson (2017), An introduction to recursive partitioning using the rpart routines, *Mayo Foundation*.

[8] Achim Zeileis, Torsten Hothorn, and Kurt Hornik (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17:2, 492 - 514, DOI: 10.1198/106186008X319331

[9] Brett Lantz (2013). *Machine Learning with R*. Livery Place, Birmingham. pg. 293 - 324.

[10] Juhani Kahri, Aino Soro Paavonen, Christian Ehnholm, and Marja Riitta Taskinen (2016); ApoE Polymorphism Is Associated With C Reactive Protein in Low HDL Family Members and in Normolipidemic Subjects, *Hindawi Publishing Corporation, Mediators of Inflammation*, Volume 2006, Article ID 12587, Pages 1-5, DOI 10.1155/MI/2006/12587.

[11] *Universidad del País Vasco* Fecha consulta: noviembre, 2017 desde https://ocw.ehu.eus/ pluginfile.php/3145/mod_resource/content/1/estadistica/tema-12-seleccion-de-modelos.pdf

[12] Ferre Jaén, M.E., *Universidad de Murcia* Fecha consulta: noviembre, 2017 desde http://gauss.inf.um.es/feir/45/

[13] Zou, H. and Hastie, T. (2004), Regularization and variable selection via the elastic net, *J. R. Statist. Soc. B*, 67, Part 2, pp. 301?320.

[14] Hastie, T and Qian J., (septiembre, 2016), Glmnet Vignette, *Universidad de Stanford*, Fecha de consulta: septiembre, 2017 desde https://web.stanford.edu/ hastie/glmnet/glmnet_alpha.html

Anexos

Anexo A: Información de las variables

Vieja codificación	Nueva codificación	Descripción de la variable	Unidades	Tipo de dato	Formato	Valores que puede adoptar	Valores normales
SEQN	**SEQN**	Número anónimo procedente de algoritmo de anonimización custodiado	NA	Numérico	Secuencia de longitud variable	Cualquiera	
DTANALIT	**DTANALIT**	Fecha de la introducción de la petición de analítica en MODULAB	NA	Fecha	Dd/mm/yyyy	min: 01/01/2009 No informado	
LBXSTB	**BilirrubinaTotal**	Bilirrubina total	mg/dL	Numérico	NN,N	min:0,1; max:70,0 NA Null No informado	0.1 - 1.2
LBXSCA	**Calcio**	Calcio	mg/dL	Numérico	N,N	min:0,8; max:30,0 NA Null No informado	120 - 220
LBXTC	**Colesterol**	Colesterol	mg/dL	Numérico	NNN	min:5; max:1090 No informado	120 - 220
LBXHDD	**HDL.Colesterol**	HDL-Colesterol	mg/dL	Numérico	NNN	min:2; max:225 No informado	40 - 60
LBXSCR	**Creatinina**	Creatinina	mg/dL	Numérico	NN,NN	min:0,20; max:30,00 No informado	07-1.3
LBXSGTSI	**GGT**	GGT	UI/L	Numérico	NNNN	min:7; max:1000 No informado	0 - 55
LBXGLU	**Glucosa**	Glucosa	mg/dL	Numérico	NNN	min:40; max:512 No informado	74 - 106
LBXSASSI	**AST.GOT**	Enzimas AST/GOT	UI/L	Numérico	NNN	min:7; max:800 No informado	0 - 50
LBXSATSI	**ALT.GPT**	Enzimas ALT/GPT	UI/L	Numérico	NNN	min:7; max:800 No informado	0 - 50
LBXTR	**Triglicéridos**	Triglicéridos	mg/dL	Numérico	NNNN	min:2; max:5000 No informado	30 - 175
LBXSUA	**AcidoUrico**	Ácido úrico	mg/dL	Numérico	NN,N	min:0,1; max:37 No informado	3.5 - 7.2
LBXAPA1	**ApolipoproteinaA1**	Apolipoproteína A1	mg/dL	Numérico	NNN	min:25; max:450 OOR; valor fuera del rango de la técnica Null; no corresponde hacer la prueba. No informado; No se dispone del dato NA; técnica en pruebas	90 - 170
LBIAPA1	**ApolipoproteinaA1.comment**	Apolipoproteína A1, comment	NA	Texto	N	1 - Fuera de rango inferior 2 - Fuera de rango superior null - No se requiere el dato NA - técnica en proceso de puesta a punto No informado	
LBXAPB	**ApolipoproteinaB**	Apolipoproteína B	mg/dL	Numérico	NNN	min:35; max:450 OOR; valor fuera del rango de la técnica Null; no corresponde hacer la prueba. No informado; No se dispone del dato NA; técnica en pruebas	56 - 162
LBIAPB	**ApolipoproteinaB. Commnet**	ApoB, comment	NA	Texto	N	1 - Fuera de rango inferior 2 - Fuera de rango superior null - No se requiere el dato NA - técnica en proceso de puesta a punto No informado	
LBXLPPA	**Lipoproteina.a**	Lipoproteína(a)	mg/dL	Numérico	NNN	min:2; max:640 OOR; valor fuera del rango de la técnica Null; no corresponde hacer la prueba. No informado; No se dispone del dato NA; técnica en pruebas	0-30
LBILPPA	**Lipoproteina.a. Comment**	Lp(a), comment	NA	Texto	N	1 - Fuera de rango inferior 2 - Fuera de rango superior null - No se requiere el dato NA; técnica en proceso de puesta a punto No informado	
LBXCRP	**ProteinaCReactiva**	Proteína C Reactiva	mg/dL	Numérico	NN,NN	min:0,02; max:144,00 OOR; valor fuera del rango de la técnica Null; no corresponde hacer la prueba. No informado; No se dispone del dato NA; técnica en pruebas	0.00 -0.50
LBICRP	**ProteinaCReactiva. Commnet**	Proteína C Reactiva, comment	NA	Texto	N	1 - Fuera de rango inferior 2 - Fuera de rango superior null - No se requiere el dato NA; técnica en proceso de puesta a punto No informado	
LBXT4F	**T4libre**	T4 libre	ng/dL	Numérico	NN,NN	min:0,25; max:6 OOR; valor fuera del rango de la técnica Null; no corresponde hacer la prueba. No informado; No se dispone del dato NA; técnica en pruebas	1.2 - 2.2
LBIT4F	**T4libre. Commnet**	T4 libre, comment	NA	Texto	N	1 - Fuera de rango inferior 2 - Fuera de rango superior null - No se requiere el dato NA; técnica en proceso de puesta a punto No informado	

Vieja codificación	Nueva codificación	Varible description	Unidades	Tipo de dato	Formato	Valores que puede adoptar	Valores normales
LBXTSH1	**TSH**	TSH	?U/mL	Numérico	NN,NN	min:0,01 max:100,00 OOR; valor fuera del rango de la técnica Null; no corresponde hacer la prueba. No informado; No se dispone del dato NA; técnica en pruebas	0.4 - 4.0
LBITSH1	**TSH. Comment**	TSH, comment	NA	Texto	N	1 - Fuera de rango inferior 2 - Fuera de rango superior null - No se requiere el dato NA; técnica en proceso de puesta a punto No informado	
LBXGH	**HemoglobinaGlicada**	Hemoglobina glicada/glicosilada	%	Numérico	NN,N	min:4,4; max:16,4 NA Null No informado	4.6 - 5.8
URXUCR	**CreatininaEnOrina**	Creatinina en orina	mg/dL	Numérico	NNN,N	min:10,0; max: 600,0 NA Null No informado	>100
URXUMS	**Microalbumina**	Microalbúmina	mg/L	Numérico	NNNN,N	min:2,0; max:8640,0 OOR; valor fuera del rango de la técnica Null; no corresponde hacer la prueba. No informado; No se dispone del dato NA; técnica en pruebas	<30
URIUMS	**Microalbumina. Commnet**	Microalbumin, comment	NA	Texto	N	1 - Fuera de rango inferior 2 - Fuera de rango superior null - No se requiere el dato NA; técnica en proceso de puesta a punto No informado	
EXXDT	dtCD	Fecha de paso del trabajador por el circuito de reconocimientos (Fecha de obtención de los DC)	días/meses/ a?os	Fecha	dd/mm/yyyy	min: 26/02/2009; No legible; No informado	
BMXWT	PesoTrabajador	Peso del trabajador	kg	Numérico	NNN,DD	min: 0;max;300; No legible; No informado	
BMXWAIST	PerimetroAbdominal	Abdominal perimeter	cm	Numérico	NNN,DD	min: 0;max;250; No legible; No informado	
HMUEST	HoraMuestra	Hora de la toma de tensión arterial y pulso cardiaco.	Horas/min	Numérico	Hora (hh24:mi)	Hora; No legible; No informado	
BPDSYAV	PresionSitolica	Presión sistólica media entre las tres medidas tomadas	mmHg	Numérico	NNN	N. entero; No legible; No informado	
BPDDIAV	PresionDiastolica	Presión diastólica media entre las tres medidas tomadas.	mmHg	Numérico	NNN	N.entero; No legible; No informado	
BPDPLAV	PulsacionesPorMinuto	60 sec. pulse average	pulsaciones/minuto	Numérico	NNN	N. entero; No legible; No informado	
ALDQWK	ConsumoAlcohol	Consumo de alcohol: cantidad	gr/week	Texto	NNNN	N. entero; No legible; No informado	
MEQDIAB	MedicaDiabetes	Toma medicación para diabetes	NA	Texto	CC	No; Si; No legible; No informado	
MEQLIPID	MedicaDislipemia	Toma medicación para dislipemia	NA	Texto	CC	No; Si; No legible; No informado	
MEQBLPR	MedicaHipertension	Toma medicación para hipertensión	NA	Texto	CC	No; Si; No legible; No informado	
MEQAPLT	MedicaAntiagregante	Toma medicación antiagregante	NA	Texto	CC	No; Si; No legible; No informado	
snp112	snp112	Marcador 112	NA	Texto	CC	CC - Citosina-Citosina CT - Citosina-Timina TT - Timina-Timina	
snp158	snp158	Marcador 158	NA	Texto	CC	CC - Citosina-Citosina CT - Citosina-Timina TT - Timina-Timina	
APOE	APOE	Identificador tipo Apolipoproteina	NA	Texto	CC	E2 - Apolipoproteína E2 E3 - Apolipoproteína E3 E4 - Apolipoproteína E4	
metSindrom.Recode	metSindrom.Recode	Identificador de si parece el sindrome	NA	Texto	CC	No; Si; No legible; No informado	

Tabla A.1: Cuadro informativo de las variables presentes en los conjuntos de datos.

A.1. Resumen numérico variables

Variable	Media	Desviación	se(mean)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
BilirrubinaTotal	0,646	0,333	0,005	0,390	0,515	2,599	11,752	0,13	0,4	0,6	0,79	3,7	3841	0
Calcio	9,390	0,407	0,007	0,600	0,043	0,630	7,558	8,1	9,1	9,4	9,7	14,7	3841	0
Colesterol	212,411	37,532	0,606	50,000	0,177	0,194	0,097	99	186	212	236	370	3841	0
HDL.Colesterol	52,561	10,958	0,177	14,000	0,208	0,726	1,393	3	45	51	59	111	3841	0
Creatinina	0,985	0,150	0,002	0,150	0,152	9,705	259,089	0,53	0,9	0,98	1,05	5,55	3841	0
GGT	43,563	34,146	0,551	21,000	0,784	6,015	66,169	12	27	34	48	697	3841	0
Glucosa	99,038	18,769	0,303	16,000	0,190	3,811	31,571	39	89	96	105	372	3841	0
AST.GOT	24,931	10,967	0,177	7,000	0,440	19,804	739,954	11	20	23	27	473	3841	0
ALT.GPT	27,981	13,361	0,216	12,000	0,478	2,572	12,247	3	20	25	32	154	3841	0
Trigliceridos	145,968	98,929	1,596	92,000	0,678	3,398	24,951	14	84	120	176	1654	3841	0
AcidoUrico	5,535	1,201	0,019	1,500	0,217	0,391	0,290	1,6	4,7	5,5	6,2	10,3	3841	0
ApolipoproteinaA1	142,495	18,989	0,307	22,000	0,133	0,562	2,194	39	131	141	153	267	3831	10
ApolipoproteinaB	106,081	25,228	0,407	34,000	0,238	0,259	0,037	36	89	105	123	215	3837	4
Lipoproteina.a	29,318	30,636	0,559	34,000	1,045	2,425	9,861	2	8	18	42	340	3007	834
ProteinaCReactiva	0,334	0,539	0,009	0,300	1,616	15,823	456,749	0,02	0,1	0,2	0,4	19,1	3485	356
T4libre	0,837	0,119	0,002	0,140	0,142	0,979	8,004	0,28	0,76	0,82	0,9	2,2	3841	0
TSH	1,659	1,617	0,026	0,950	0,975	19,190	583,544	0,03	1,01	1,425	1,96	60,31	3840	1
HemoglobinaGlicada	5,498	0,583	0,009	0,400	0,106	4,740	36,661	3,6	5,2	5,4	5,6	12,9	3840	1
CreatininaEnOrina	163,061	65,587	1,103	87,250	0,402	0,620	0,570	1,7	115,75	156,8	203	494	3535	306
Microalbumina	11,059	38,705	0,671	4,880	3,500	17,722	412,424	2	3,5	5,3	8,38	1120	3331	510
PesoTrabajador	81,517	11,573	0,189	15,150	0,142	0,505	0,703	48,8	73,45	80,5	88,6	146,1	3759	82
PerimetroAbdominal	97,069	9,771	0,159	12,500	0,101	0,267	0,340	66	90,5	96,7	103	135,8	3765	76
PresionSitolica	126,795	14,491	0,237	17,000	0,114	0,930	1,847	88	117	125	134	220	3753	88
PresionDiastolica	83,787	9,845	0,161	13,000	0,117	0,368	0,661	52	77	84	90	130	3753	88
PulsacionesPorMinuto	71,059	11,829	0,193	15,000	0,166	0,541	0,559	38	63	70	78	128	3753	88
ConsumoAlcohol	66,670	59,667	0,980	100,000	0,895	0,831	2,250	0	0	90	100	500	3707	134

Tabla A.2: Resumen numérico de las variables.

A.1.1. Por grupos

Variables numéricas

Resúmenes numéricos por grupos de apolipoproteínas. Recordemos que el objetivo de estos resúmenes es dar una visión global del conjunto de datos y a su vez identificar posibles variables influyentes en la determinación de los grupos. A continuación mostramos cada uno de los resúmenes:

Variable:	BilirrubinaTotal													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	0,624	0,334	0,017	0,300	0,536	2,724	12,891	0,2	0,4	0,52	0,7	3,1	388	0
E3	0,650	0,328	0,006	0,400	0,504	2,485	10,577	0,13	0,4	0,6	0,8	3,5	2694	0
E4	0,646	0,346	0,013	0,308	0,536	2,772	13,667	0,2	0,4	0,6	0,708	3,7	670	0

Variable:	Calcio													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	9,388	0,406	0,021	0,600	0,043	-0,263	-0,182	8,3	9,1	9,4	9,7	10,5	388	0
E3	9,388	0,397	0,008	0,500	0,042	0,135	0,463	8,1	9,1	9,4	9,6	11,3	2694	0
E4	9,386	0,387	0,015	0,500	0,041	-0,021	-0,177	8,2	9,1	9,4	9,6	10,4	670	0

Variable:	Colesterol													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	200,593	37,232	1,890	51,250	0,186	0,172	-0,065	117	175,75	201,5	227	338	388	0
E3	213,094	37,197	0,717	50,000	0,175	0,228	0,162	106	187	212	237	370	2694	0
E4	216,778	37,694	1,456	48,750	0,174	0,077	-0,044	99	192,25	217	241	336	670	0

Variable:	HDL,Colesterol													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	52,884	10,862	0,551	12,250	0,205	0,847	1,735	28	45,75	52	58	106	388	0
E3	52,944	11,049	0,213	14,000	0,209	0,744	1,333	13	45	52	59	111	2694	0
E4	51,272	10,674	0,412	13,000	0,208	0,594	1,453	3	44	50	57	96	670	0

Variable:	Creatinina													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	0,9813	0,1191	0,0060	0,1500	0,1213	0,4280	0,3836	0,71	0,9	0,97	1,05	1,41	388	0
E3	0,9857	0,1606	0,0031	0,1500	0,1629	11,1105	279,7434	0,53	0,9	0,98	1,05	5,55	2694	0
E4	0,9839	0,1225	0,0047	0,1475	0,1245	0,9799	5,5450	0,63	0,903	0,98	1,05	1,93	670	0

Tabla A.3: Resumen numérico por grupos (1).

Variable:	GGT													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	42,216	25,709	1,305	21,000	0,609	2,837	10,324	15	27	34	48	191	388	0
E3	43,327	34,125	0,657	21,000	0,788	6,507	78,977	14	27	34	48	697	2694	0
E4	44,233	36,329	1,404	21,000	0,821	4,905	35,064	14	27	34	48	438	670	0

Variable:	Glucosa													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	97,938	17,519	0,889	14,250	0,179	4,051	33,280	53	89	96	103,25	277	388	0
E3	98,856	18,078	0,348	16,000	0,183	3,711	32,221	39	89	96	105	372	2694	0
E4	99,506	18,606	0,719	18,000	0,187	2,554	12,016	60	88	97	106	243	670	0

Variable:	AST.GOT													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	25,825	24,204	1,229	7,000	0,937	16,537	303,205	13	20	23	27	473	388	0
E3	24,704	7,776	0,150	7,000	0,315	4,123	35,512	11	20	23	27	127	2694	0
E4	25,300	9,933	0,384	7,000	0,393	6,352	68,260	14	20	23	27	162	670	0

Variable:	ALT.GPT													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	27,889	14,644	0,743	13,250	0,525	3,535	21,222	9	19	24	32,25	147	388	0
E3	27,906	12,963	0,250	12,000	0,465	2,448	11,272	6	20	25	32	154	2694	0
E4	27,957	13,535	0,523	14,000	0,484	2,468	10,917	3	19	25	33	140	670	0

Variable:	Trigliceridos													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	158,206	91,188	4,629	105,250	0,576	1,337	2,026	14	94	135	199,25	550	388	0
E3	141,513	96,560	1,860	88,000	0,682	4,052	35,678	27	82	117	170	1654	2694	0
E4	152,802	106,651	4,120	93,750	0,698	2,321	6,891	34	86,25	119	180	763	670	0

Tabla A.4: Resumen numérico por grupos (2).

Variable:	AcidoUrico													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	5,480	1,203	0,061	1,625	0,219	0,653	0,708	2,7	4,6	5,3	6,23	10,3	388	0
E3	5,527	1,194	0,023	1,500	0,216	0,416	0,383	1,6	4,7	5,4	6,2	10,1	2694	0
E4	5,571	1,226	0,047	1,600	0,220	0,208	-0,118	2,1	4,7	5,5	6,3	9,5	670	0

Variable:	ApolipoproteinaA1													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	145,518	18,529	0,941	22,000	0,127	0,519	0,853	100	134	143,5	156	214	388	0
E3	142,753	19,180	0,370	23,000	0,134	0,616	2,739	39	131	141	154	267	2685	9
E4	140,025	18,260	0,706	22,000	0,130	0,405	0,607	79	129	138	151	209	669	1

Variable:	ApolipoproteinaB													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	93,734	23,385	1,189	32,000	0,249	0,134	-0,019	41	78	94	110	181	387	1
E3	106,574	25,154	0,485	34,000	0,236	0,284	-0,032	36	89	105	123	215	2692	2
E4	110,926	24,336	0,941	31,600	0,219	0,223	0,155	39	94,4	110	126	211	669	1

Variable:	Lipoproteina,a													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	27,806	30,622	1,756	26,500	1,101	2,443	7,940	2	8	17	34,5	199	304	84
E3	29,165	30,670	0,668	33,600	1,052	2,583	11,635	2	8,4	18	42	340	2107	587
E4	31,433	31,297	1,358	38,000	0,996	1,840	4,583	2	8	19,3	46	209	531	139

Variable:	ProteinaCReactiva													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	0,337	0,385	0,020	0,300	1,141	4,132	23,300	0,02	0,1	0,2	0,4	3,1	361	27
E3	0,338	0,556	0,011	0,300	1,645	17,754	540,633	0,02	0,1	0,2	0,4	19,1	2461	233
E4	0,290	0,467	0,019	0,200	1,609	9,623	135,061	0,03	0,1	0,2	0,3	7,9	580	90

Tabla A.5: Resumen numérico por grupos (3).

Variable:	T4libre													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	0,837	0,109	0,006	0,130	0,130	0,483	0,646	0,57	0,77	0,83	0,9	1,21	388	0
E3	0,837	0,119	0,002	0,140	0,142	0,818	6,885	0,28	0,76	0,82	0,9	2,2	2694	0
E4	0,833	0,120	0,005	0,140	0,144	1,961	16,708	0,52	0,76	0,82	0,9	2,06	670	0

Variable:	TSH													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	1,587	0,992	0,050	0,895	0,625	4,240	31,196	0,32	1,008	1,36	1,9025	10,94	388	0
E3	1,684	1,840	0,035	0,960	1,092	18,422	495,413	0,05	1,01	1,43	1,97	60,31	2693	1
E4	1,585	0,802	0,031	0,880	0,506	1,515	4,526	0,03	1,04	1,445	1,92	6,9	670	0

Variable:	HemoglobinaGlicada													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	5,483	0,502	0,025	0,425	0,092	3,802	25,908	4,1	5,2	5,4	5,625	9,6	388	0
E3	5,496	0,585	0,011	0,400	0,107	5,071	41,477	3,6	5,2	5,4	5,6	12,9	2694	0
E4	5,497	0,583	0,023	0,400	0,106	4,413	30,051	4,1	5,2	5,4	5,6	10,9	669	1

Variable:	CreatininaEnOrina													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	159,342	65,859	3,515	76,750	0,413	0,720	0,771	14,3	115,5	151,9	192,25	396,1	351	37
E3	163,180	65,243	1,310	88,400	0,400	0,585	0,430	1,7	115,9	156,6	204,3	444,3	2481	213
E4	164,020	67,454	2,711	89,250	0,411	0,712	0,976	18,2	113,95	159,2	203,2	494	619	51

Variable:	Microalbumina													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	11,558	39,198	2,151	4,458	3,391	12,930	196,634	2	3,368	5	7,825	635	332	56
E3	10,542	35,226	0,728	4,700	3,342	19,025	493,309	2	3,6	5,3	8,3	1120	2343	351
E4	12,610	50,305	2,098	5,000	3,989	16,172	311,641	2	3,5	5,4	8,5	1040	575	95

Tabla A.6: Resumen numérico por grupos (4).

Variable:	PesoTrabajador													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	82,074	12,063	0,612	14,675	0,147	0,825	2,071	53	74	81,05	88,675	146,1	388	0
E3	81,489	11,500	0,222	15,200	0,141	0,457	0,528	48,8	73,4	80,5	88,6	137	2688	6
E4	81,163	11,422	0,441	15,350	0,141	0,445	0,423	52,4	73,15	80,45	88,5	126,4	670	0

Variable:	PerimetroAbdominal													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	97,530	10,121	0,514	13,350	0,104	0,248	0,555	68,8	90,675	97,3	104,025	135,8	388	0
E3	97,007	9,767	0,188	12,800	0,101	0,261	0,282	66	90,3	96,6	103,1	133,2	2694	0
E4	96,926	9,505	0,367	11,750	0,098	0,274	0,444	68,2	90,825	96,5	102,575	132	670	0

Variable:	PresionSitolica													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	125,838	13,812	0,701	18,000	0,110	0,716	0,840	88	116	124	134	180	388	**0**
E3	126,659	14,165	0,273	18,000	0,112	0,800	1,265	90	117	125	135	197	2694	**0**
E4	127,839	15,988	0,618	18,000	0,125	1,323	3,280	93	117	126	135	220	670	**0**

Variable:	PresionDiastolica													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	83,691	9,818	0,498	13,000	0,117	0,280	0,354	59	77	84	90	122	388	0
E3	83,722	9,766	0,188	13,000	0,117	0,289	0,430	55	77	84	90	130	2694	0
E4	84,082	10,164	0,393	13,000	0,121	0,692	1,585	52	77	83	90	129	670	0

Variable:	PulsacionesPorMinuto													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	71,387	11,509	0,584	15,000	0,161	0,549	0,372	42	63	70	78	113	388	0
E3	71,039	11,831	0,228	15,000	0,167	0,484	0,514	38	63	70	78	121	2694	0
E4	70,930	12,016	0,464	14,000	0,169	0,769	0,855	40	63	69	77	128	670	0

Variable:	ConsumoAlcohol													
APOE	Media	Desviación	se(Media)	IQR	cv	skewness	kurtosis	0 %	25 %	50 %	75 %	100 %	n	NA
E2	66,749	58,486	2,989	100,000	0,876	0,848	2,203	0	0	80	100	360	383	5
E3	66,406	59,030	1,146	100,000	0,889	0,719	1,596	0	0	80	100	500	2654	40
E4	67,646	62,858	2,449	100,000	0,929	1,199	4,322	0	0	100	100	450	659	11

Tabla A.7: Resumen numérico por grupos (5).

Variables categóricas

Tablas de contingencia para cada una de las variables categóricas, con resúmenes numéricos por columnas y por filas y con un test de independencia de Pearson.

Frequency table:	MedicaAntiagregante	
APOE	Si	No
E2	13	375
E3	68	2624
E4	20	650

Row percentages:	MedicaAntiagregante			
APOE	Si	No	Total	Count
E2	3.4	96.6	100	388
E3	2.5	97.5	100	2692
E4	3.0	97.0	100	670

Column percentages:	MedicaAntiagregante	
APOE	Si	No
E2	12.9	10.3
E3	67.3	71.9
E4	19.8	17.8
Total	100.0	100.0
Count	101.0	3649.0

Pearson's Chi-squared test					
X-squared =	11.446,00	df =	2	p-value =	0.5642

Frequency table:	MedicaDiabetes	
APOE	Si	No
E2	16	368
E3	87	2593
E4	26	643

Row percentages:	MedicaDiabetes			
APOE	Si	No	Total	Count
E2	4.2	95.8	100	384
E3	3.2	96.8	100	2680
E4	3.9	96.1	100	669

Column percentages:	MedicaDiabetes	
APOE	Si	No
E2	12.4	10.2
E3	67.4	71.9
E4	20.2	17.8
Total	100.0	99.9
Count	129.0	3604.0

Pearson's Chi-squared test					
X-squared =	13.061,00	df =	2	p-value =	0.5205

Tabla A.8: Tablas de contingencia para las variables *MedicaAntiagregante* y *MedicaDiabetes*.

Frequency table:	MedicaDislipemia	
APOE	Si	No
E2	25	354
E3	247	2405
E4	76	583

Row percentages:	MedicaDislipemia			
APOE	Si	No	Total	Count
E2	6.6	93.4	100	379
E3	9.3	90.7	100	2652
E4	11.5	88.5	100	659

Column percentages:	MedicaDislipemia	
APOE	Si	No
E2	7.2	10.6
E3	71.0	72.0
E4	21.8	17.4
Total	100.0	100.0
Count	348.0	3342.0

Pearson's Chi-squared test					
X-squared =	70.159,00	df =	2	p-value =	0.02996

Frequency table:	MedicaHipertension	
APOE	Si	No
E2	54	334
E3	465	2229
E4	123	547

Row percentages:	MedicaHipertension			
APOE	Si	No	Total	Count
E2	13.9	86.1	100	388
E3	17.3	82.7	100	2694
E4	18.4	81.6	100	670

Column percentages:	MedicaHipertension	
APOE	Si	No
E2	8.4	10.7
E3	72.4	71.7
E4	19.2	17.6
Total	100.0	100.0
Count	642.0	3110.0

Pearson's Chi-squared test					
X-squared =	35.672,00	df =	2	p-value =	0.168

Tabla A.9: Tablas de contingencia para las variables *MedicaDislipemia* y *MedicaHipertension*.

Frequency table:	metSindrom.Recode	
APOE	Si	No
E2	99	289
E3	733	1961
E4	203	467

Row percentages:	metSindrom.Recode			
APOE	Si	No	Total	Count
E2	25.5	74.5	100	388
E3	27.2	72.8	100	2694
E4	30.3	69.7	100	670

Column percentages:	metSindrom.Recode	
APOE	Si	No
E2	9.6	10.6
E3	70.8	72.2
E4	19.6	17.2
Total	100.0	100.0
Count	1035.0	2717.0

Pearson's Chi-squared test					
X-squared =	34.926,00	df =	2	p-value =	0.1744

Tabla A.10: Tabla de contingencia para la variable *metSindrom.Recode*.

A.2. Gráficos

A.2.1. Variables numéricas

Boxplot

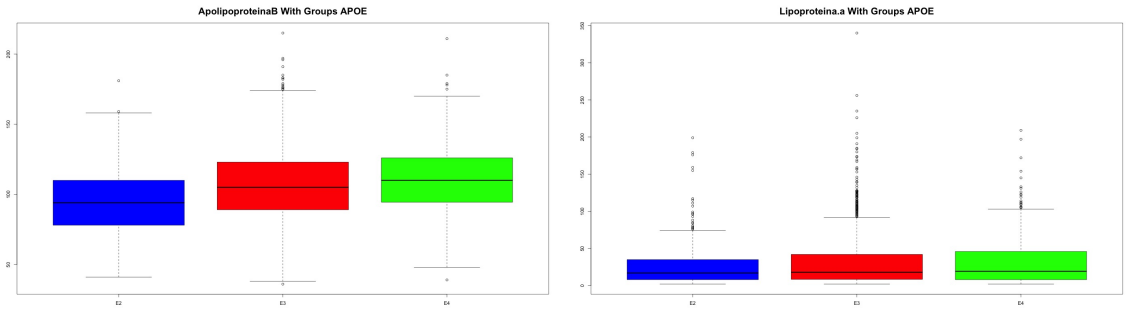

Figura A.1: Diagrama de cajas para las variables *ApolipoproteinaB* y *Lipoproteina.a*.

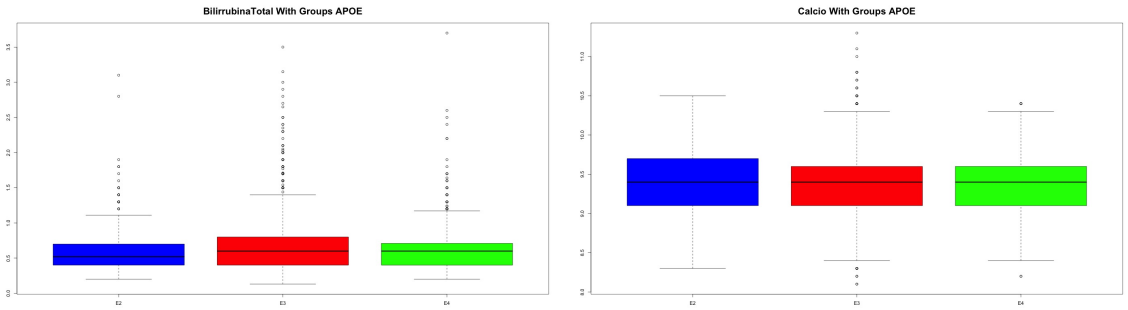

Figura A.2: Diagrama de cajas para las variables *BilirrubinaTotal* y *Calcio*.

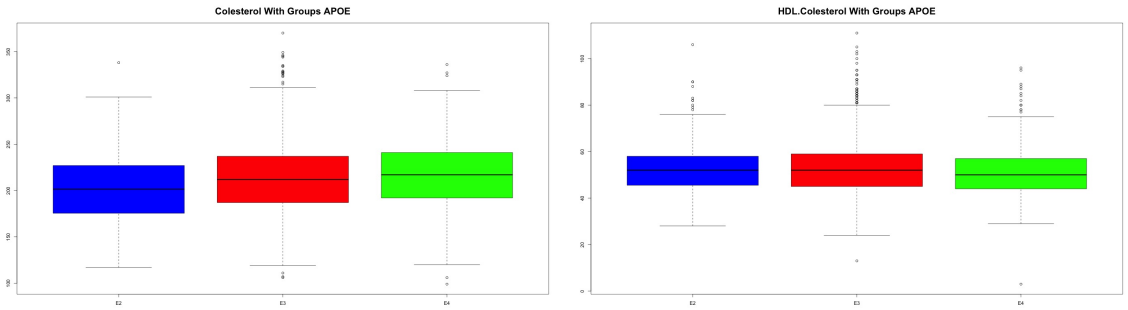

Figura A.3: Diagrama de cajas para las variables *Colesterol* y *HDL.Colesterol*.

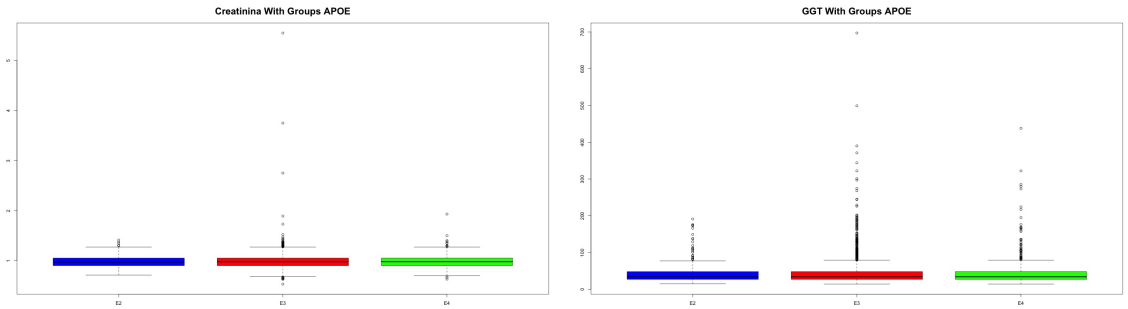

Figura A.4: Diagrama de cajas para las variables *Creatinina* y *GGT*.

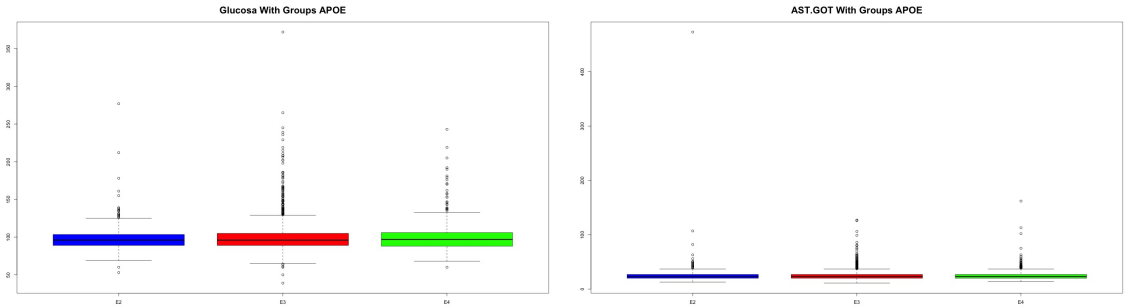

Figura A.5: Diagrama de cajas para las variables *Glucosa* y *AST.GOT*.

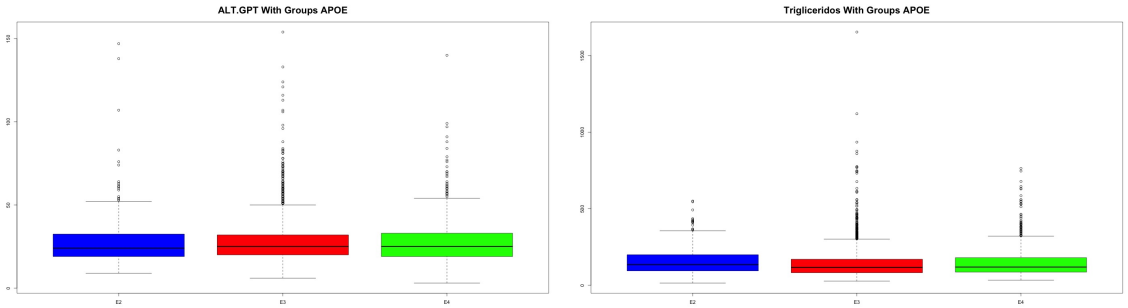

Figura A.6: Diagrama de cajas para las variables *AST.GPT* y *Triglicéridos*.

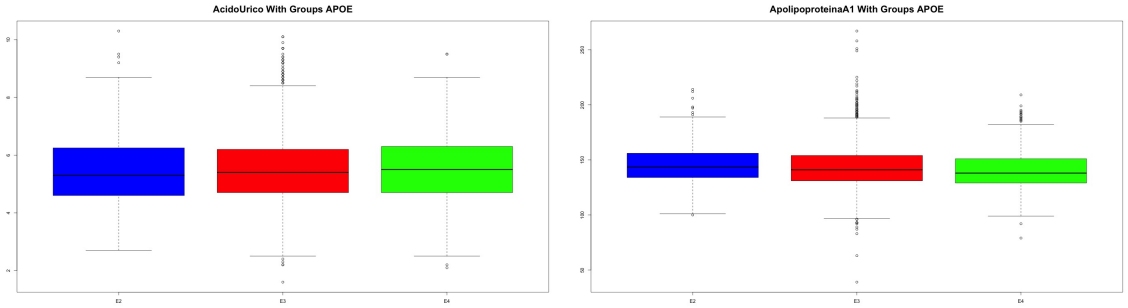

Figura A.7: Diagrama de cajas para las variables *AcidoÚrico* y *ApolipoproteinaA1*.

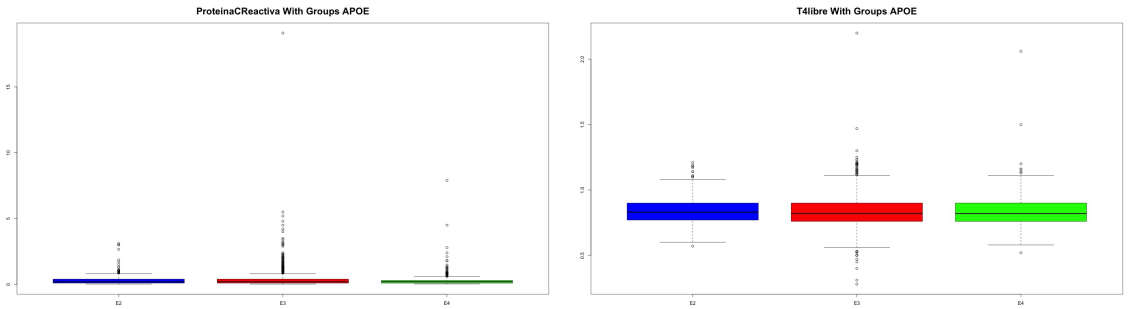

Figura A.8: Diagrama de cajas para las variables *ProteinaCReactiva* y *T4libre*.

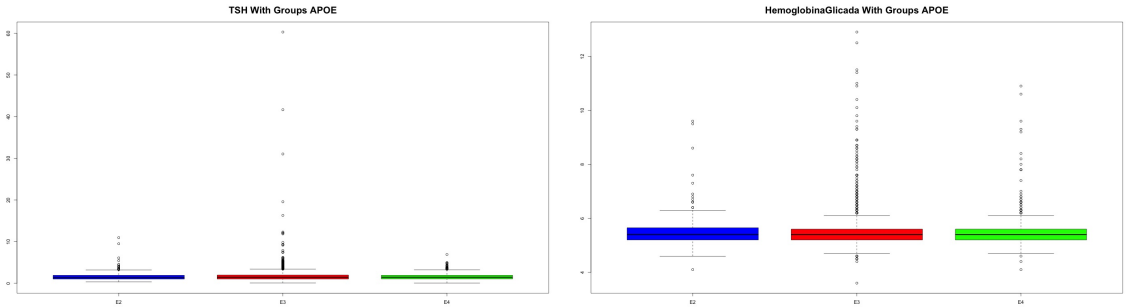

Figura A.9: Diagrama de cajas para las variables *TSH* y *HemoglobinaGlicada*.

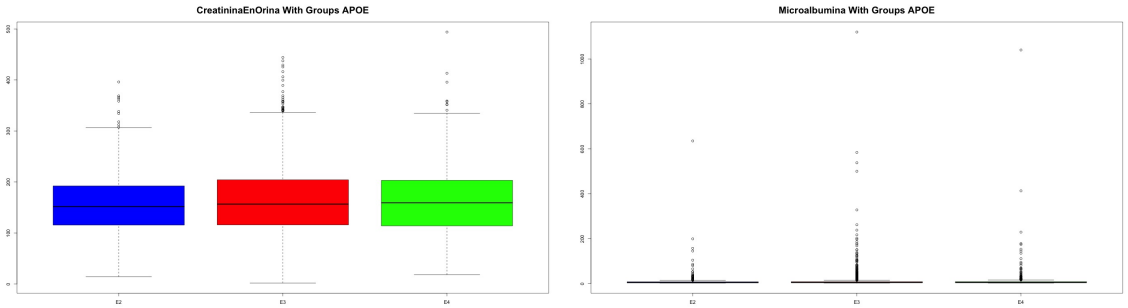

Figura A.10: Diagrama de cajas para las variables *CreatininaEnOrina* y *Microalbumina*.

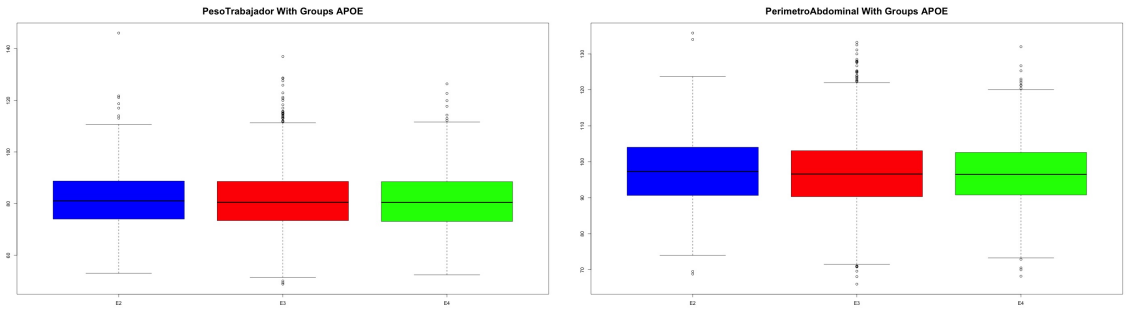

Figura A.11: Diagrama de cajas para las variables *PesoTrabajador* y *PerímetroAbdominal*.

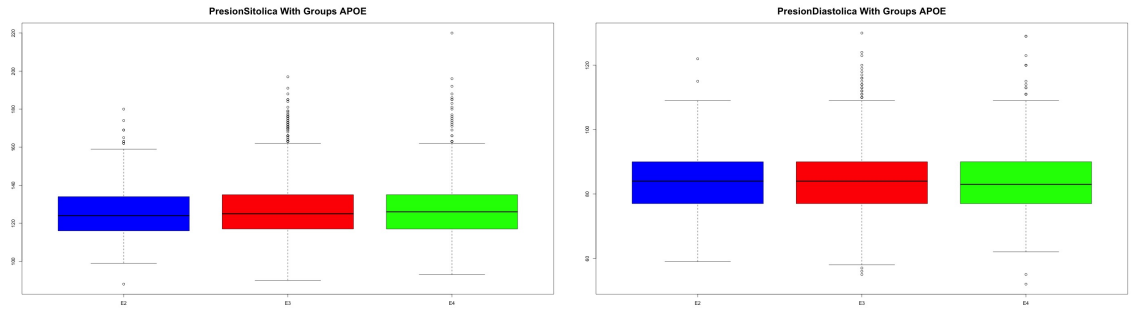

Figura A.12: Diagrama de cajas para las variables *PresionSistolica* y *PresionDiastolica*.

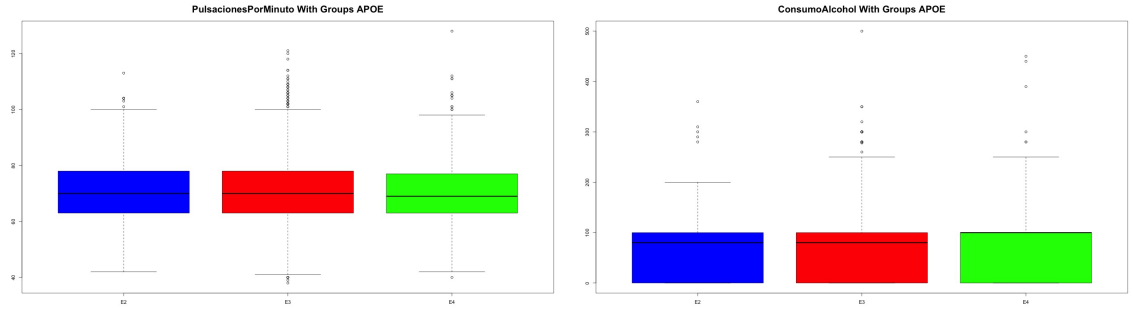

Figura A.13: Diagrama de cajas para las variables *PulsacionesPorMinuto* y *ConsumoAlcohol*.

Gráficas de densidad

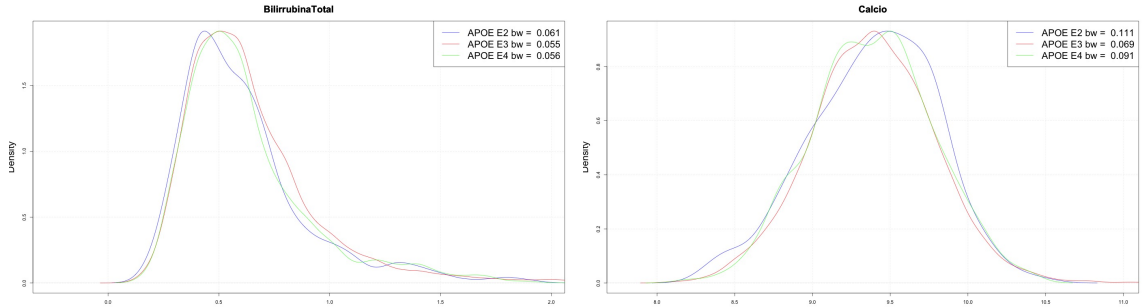

Figura A.14: Gráficas de densidad para las variables *BilirribuninaTotal* y *Calcio*.

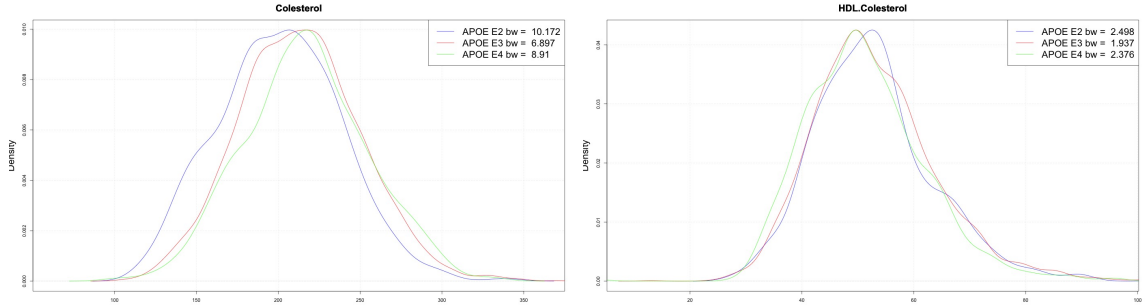

Figura A.15: Gráficas de densidad para las variables *Colesterol* y *HDL.Colesterol*.

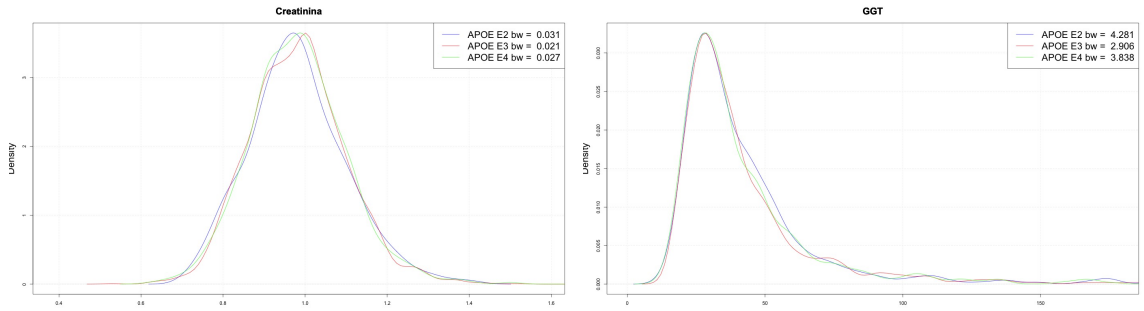

Figura A.16: Gráficas de densidad para las variables *Creatinina* y *GGT*.

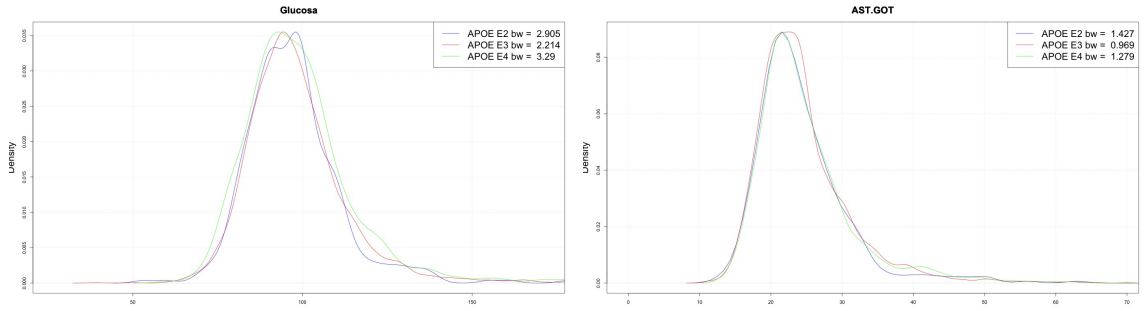

Figura A.17: Gráficas de densidad para las variables *Glucosa* y *AST.GOT*.

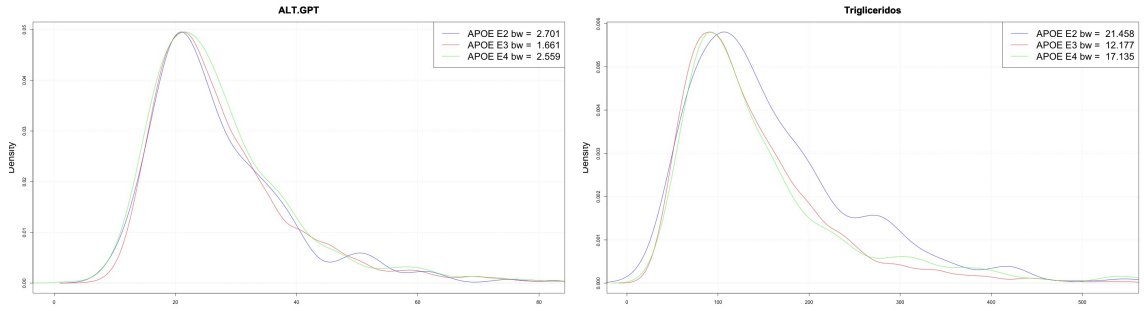

Figura A.18: Gráficas de densidad para las variables *ALT.GPT* y *Triglicéridos*.

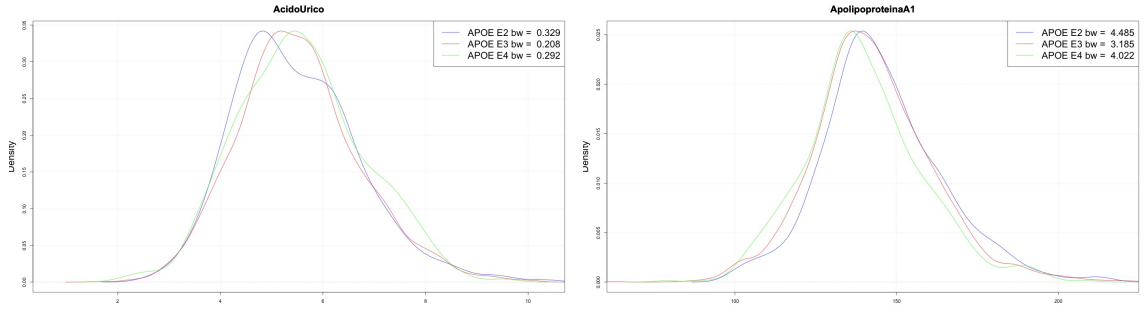

Figura A.19: Gráficas de densidad para las variables *AcidoUrico* y *ApolipoproteinaA1*.

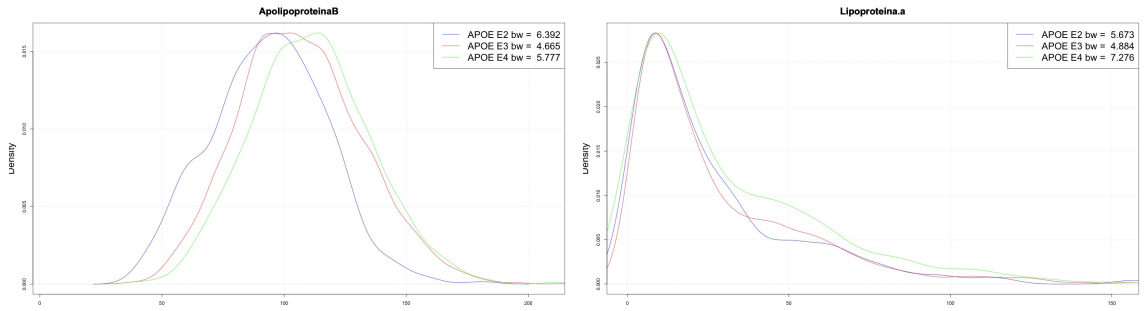

Figura A.20: Gráficas de densidad para las variables *ApolipoproteinaB* y *Lipoproteina.a*.

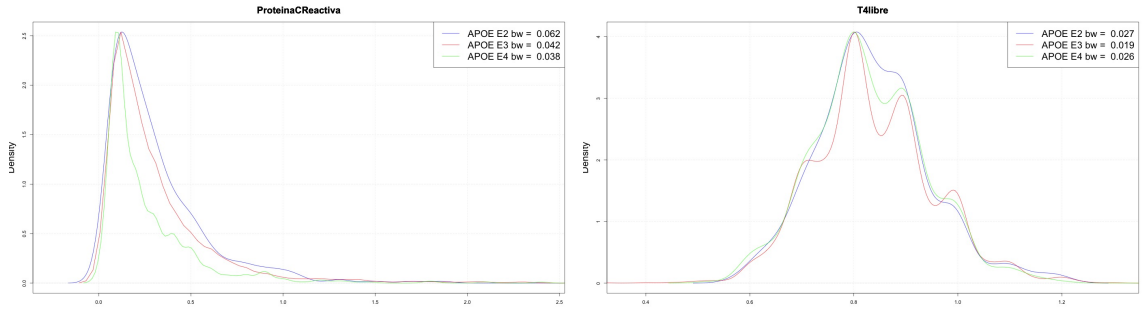

Figura A.21: Gráficas de densidad para las variables *ProteinaCReactiva* y *T4libre*.

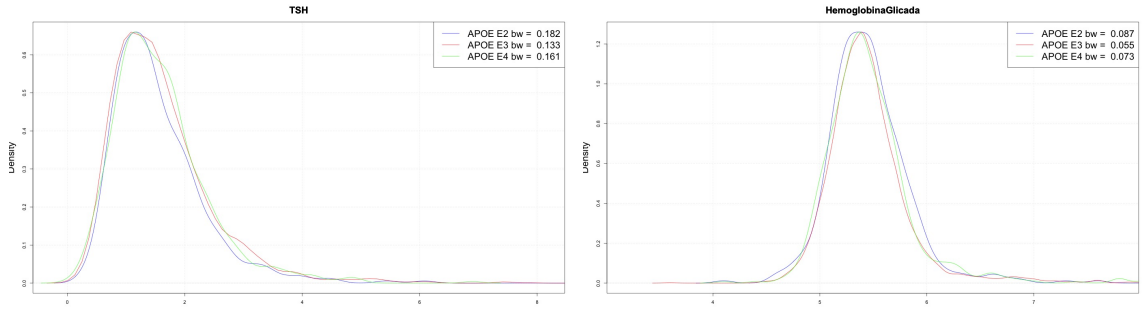

Figura A.22: Gráficas de densidad para las variables *TSH* y *HemoglobinaGlicosilada*.

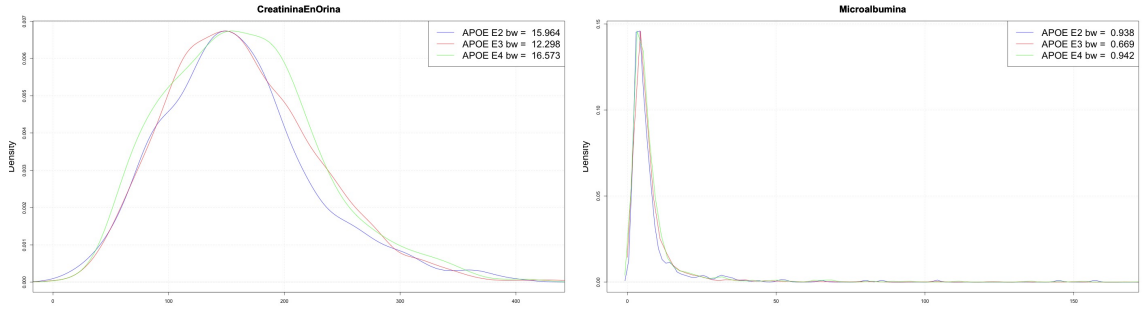

Figura A.23: Gráficas de densidad para las variables *CreatininaEnOrina* y *Microalbuminia* .

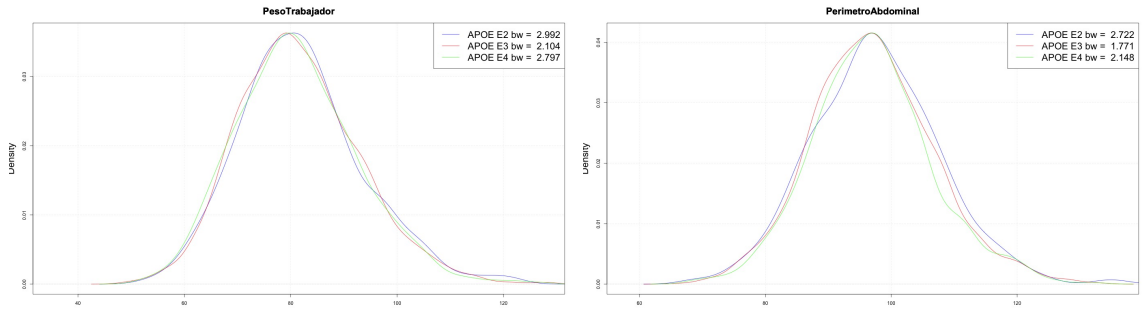

Figura A.24: Gráficas de densidad para las variables *PesoTrabajador* y *PerimetroAbdominal*.

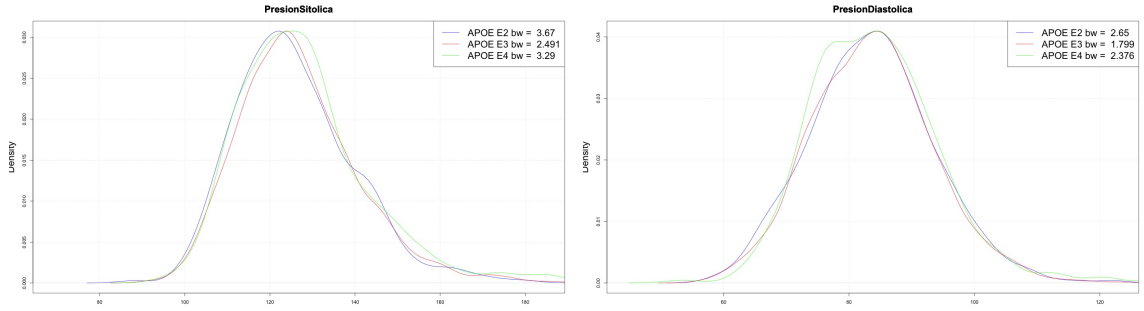

Figura A.25: Gráficas de densidad para las variables *PresionSistolica* y *PresionDiastolica*.

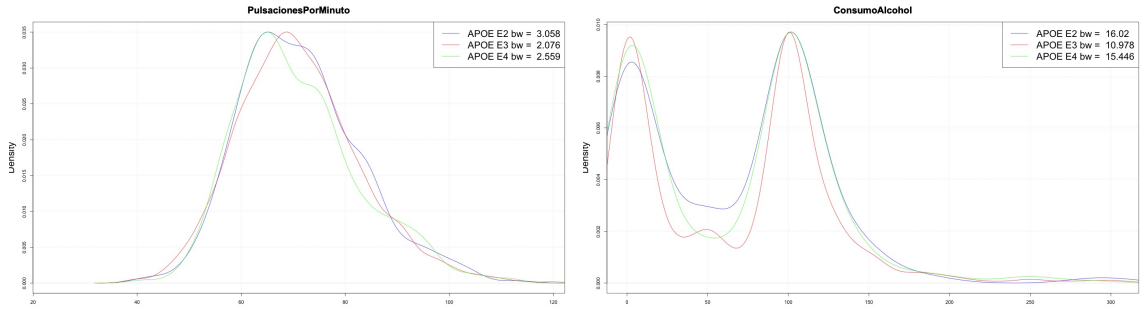

Figura A.26: Gráficas de densidad para las variables *PulsacionesPorMinuto* y *ConsumoAlcohol*.

A.2.2. Variables categóricas

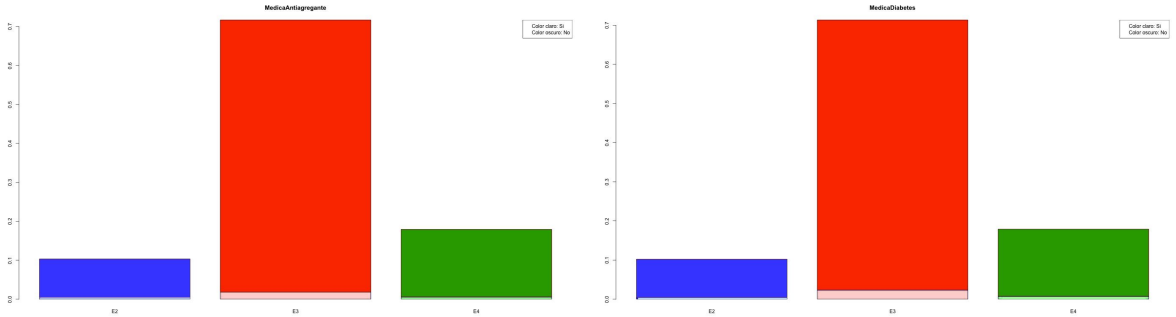

Figura A.27: Diagrama de barras de las variables *MedicaAntiagregante* y *MedicaDiabetes*.

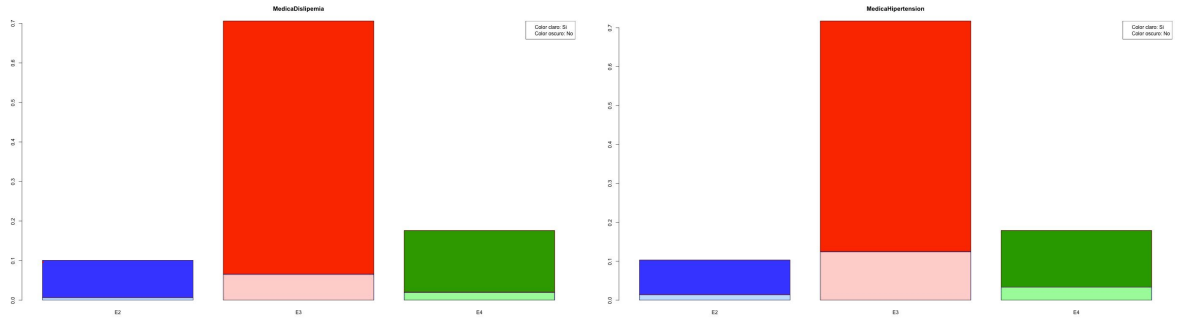

Figura A.28: Diagrama de barras de las variables *MedicaDislipemia* y *MedicaHipertension*.

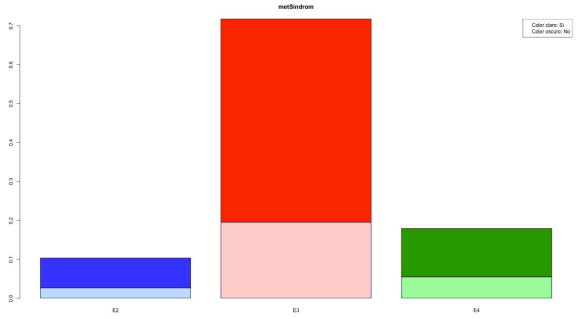

Figura A.29: Diagrama de barras de la variable *metSindrom*.

Anexo B: Modelos auxiliares

B.1. Modelos intermedios de regresión logística.

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-2.808	0.461	-6.087	0.000	***
ApoBApoA1	1.273	0.269	4.739	0.000	***
ColesterolEF[T.Alto]	-0.052	0.117	-0.446	0.655	
Lipoproteina.a	0.004	0.002	2.285	0.022	*
ProteinaCReactiva	-1.085	0.338	-3.210	0.001	**
ConsumoAlcohol.MDis	0.003	0.002	2.150	0.032	*
PresionSitolica	0.004	0.003	1.075	0.282	
Microalbumina.metSin	0.002	0.001	1.402	0.161	
ProteinaCReactivaEF[T.Alto]	0.057	0.232	0.247	0.805	

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-2.811	0.461	-6.095	0.000	***
ApoBApoA1	1.276	0.268	4.753	0.000	***
ColesterolEF[T.Alto]	-0.052	0.117	-0.448	0.654	
Lipoproteina.a	0.004	0.002	2.275	0.023	*
ProteinaCReactiva	-1.021	0.215	-4.759	0.000	***
ConsumoAlcohol.MDis	0.003	0.002	2.148	0.032	*
PresionSitolica	0.004	0.003	1.063	0.288	
Microalbumina.metSin	0.002	0.001	1.417	0.156	

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-2.782	0.456	-6.097	0.000	***
ApoBApoA1	1.216	0.232	5.236	0.000	***
Lipoproteina.a	0.004	0.002	2.240	0.025	*
ProteinaCReactiva	-1.018	0.215	-4.746	0.000	***
ConsumoAlcohol.MDis	0.003	0.002	2.152	0.031	*
PresionSitolica	0.004	0.003	1.055	0.292	
Microalbumina.metSin	0.002	0.001	1.415	0.157	

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-2.345	0.189	-12.430	0.000	***
ApoBApoA1	1.249	0.230	5.444	0.000	***
Lipoproteina.a	0.003	0.002	2.151	0.032	*
ProteinaCReactiva	-0.998	0.213	-4.684	0.000	***
ConsumoAlcohol.MDis	0.003	0.002	2.202	0.028	*
Microalbumina.metSin	0.002	0.001	1.488	0.137	

Tabla B.1: Secuencia de tablas perteneciente a cada uno de los modelos intermedios en la regresión logística (1).

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-2.356	0.189	-12.492	0.000	***
ApoBApoA1	1.269	0.229	5.536	0.000	***
Lipoproteina.a	0.003	0.002	2.128	0.033	*
ProteinaCReactiva	-0.976	0.211	-4.622	0.000	***
ConsumoAlcohol.MDis	0.004	0.002	2.302	0.021	*

Tabla B.2: Secuencia de tablas perteneciente a cada uno de los modelos intermedios en la regresión logística (2).

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-2.344	0.188	-12.447	<2E-16	***
ApoBApoA1	1.278	0.229	5.578	0.000	***
Lipoproteina.a	0.003	0.002	2.201	0.028	*
ProteinaCReactiva	-0.955	0.211	-4.525	0.000	***

Tabla B.3: Modelo final de la regresión logística.

B.2. Árboles de clasificación

Modelo con el criterio Información

```
Call:
rpart(formula = CasosE4 ~ ., data = BaseModels[, -34], method = "class",
    parms = list(split = "information"), control = rpart.control(xval = 10,
        cp = 0))
  n= 3764

             CP nsplit rel error   xerror       xstd
1  0.0033333333      0 1.0000000 1.000000 0.03486842
2  0.0029629630     26 0.8696296 1.066667 0.03574868
3  0.0024691358     43 0.8148148 1.137778 0.03662882
4  0.0022222222     52 0.7896296 1.207407 0.03743584
5  0.0019753086     60 0.7718519 1.237037 0.03776368
6  0.0014814815     66 0.7600000 1.260741 0.03801950
7  0.0012698413     85 0.7244444 1.327407 0.03870927
8  0.0012345679    112 0.6622222 1.339259 0.03882743
9  0.0011851852    118 0.6548148 1.345185 0.03888602
10 0.0009876543    136 0.6296296 1.374815 0.03917405
11 0.0007407407    141 0.6222222 1.388148 0.03930105
12 0.0003703704    143 0.6207407 1.459259 0.03995157
13 0.0000000000    147 0.6192593 1.485926 0.04018418

Variable importance
                    ICM     PerimetroAbdominal                     ICA
              84.8206583             84.3954209             80.7521076
       RatioColesterol          Trigliceridos       ApolipoproteinaB
              80.6720423             80.0477851             75.9848682
             ApoBApoA1                    GGT             Colesterol
              75.2443529             62.7746456             62.7500951
         HDL.Colesterol         PesoTrabajador             AcidoUrico
              61.0381209             52.5631319             49.1880476
```

```
              AST.GOT                ALT.GPT      CreatininaEnOrina
            46.1483259             45.4182047             44.7925892
       PresionSitolica       BilirrubinaTotal      ProteinaCReactiva
            44.3227655             41.1926138             39.1658360
     ApolipoproteinaA1                Glucosa                    TSH
            39.0864291             38.0898812             36.9590866
               Calcio HemoglobinaGlicosilada                T4libre
            33.3904972             33.2785849             33.0209009
                 Edad         Lipoproteina.a          Microalbumina
            32.9005107             29.5830767             26.9834581
     PresionDiastolica   PulsacionesPorMinuto             Creatinina
            25.4442614             23.8622246             17.8306831
        ConsumoAlcohol      metSindrom.Recode     MedicaHipertension
            10.1891475              6.4496234              1.7330106
        MedicaDiabetes    MedicaAntiagregante
             0.6134553              0.2667595
```

Después de podar

```
Call:
rpart(formula = CasosE4 ~ ., data = BaseModels[, -c(remove, 34)],
    method = "class", parms = list(split = "information"), control = rpart.control(xval = 10
        cp = 0))
  n= 3764

           CP nsplit rel error   xerror       xstd
1 0.004074074      0 1.0000000 1.000000 0.03486842
2 0.003851900      7 0.9674074 1.022222 0.03516802

Variable importance
ProteinaCReactiva  ApolipoproteinaB     Trigliceridos   RatioColesterol
       24.2305822        17.5649956        12.2419389        12.0916914
              GGT    HDL.Colesterol              Edad           T4libre
        8.3671378         6.1665379         5.2340445         4.8541841
              ICM           Glucosa               TSH        AcidoUrico
        4.7423164         4.5592313         1.2135460         1.1674178
 BilirrubinaTotal           AST.GOT    MedicaDiabetes   PresionSitolica
        0.6067730         0.5145166         0.2747437         0.1373718

Node number 1: 3764 observations,    complexity param=0.004074074
  predicted class=0  expected loss=0.1793305  P(node) =1
    class counts:  3089   675
   probabilities: 0.821 0.179
  left son=2 (2231 obs) right son=3 (1533 obs)
  Primary splits:
      ProteinaCReactiva < 0.115    to the right, improve=24.230580, (2 missing)
      ApolipoproteinaB  < 91.95    to the left,  improve=17.699450, (4 missing)
      RatioColesterol   < 3.887626 to the left,  improve=11.792990, (0 missing)
      HDL.Colesterol    < 42.5     to the right, improve= 6.771839, (0 missing)
      Trigliceridos     < 280.5    to the left,  improve= 6.545651, (0 missing)
  Surrogate splits:
      ICM             < 25.89421 to the right, agree=0.636, adj=0.106, (2 split)
      GGT             < 25.5     to the right, agree=0.635, adj=0.104, (0 split)
```

```
      Edad            < 43.5     to the right, agree=0.616, adj=0.057, (0 split)
      RatioColesterol < 3.426078 to the right, agree=0.612, adj=0.048, (0 split)
      Trigliceridos   < 80.5     to the right, agree=0.606, adj=0.033, (0 split)

Node number 2: 2231 observations
  predicted class=0  expected loss=0.1429852  P(node) =0.5927205
    class counts:  1912   319
   probabilities: 0.857 0.143

Node number 3: 1533 observations,    complexity param=0.004074074
  predicted class=0  expected loss=0.2322244  P(node) =0.4072795
    class counts:  1177   356
   probabilities: 0.768 0.232
  left son=6 (560 obs) right son=7 (973 obs)
  Primary splits:
      ApolipoproteinaB < 91.95    to the left,  improve=17.565000, (0 missing)
      RatioColesterol  < 3.887626 to the left,  improve=15.525590, (0 missing)
      Trigliceridos    < 296.5    to the left,  improve= 8.256002, (0 missing)
      HDL.Colesterol   < 43.5     to the right, improve= 8.246927, (0 missing)
      ICM              < 26.07352 to the left,  improve= 6.632332, (0 missing)
  Surrogate splits:
      RatioColesterol < 3.491646 to the left,  agree=0.810, adj=0.479, (0 split)
      Edad            < 37.5     to the left,  agree=0.715, adj=0.220, (0 split)
      Trigliceridos   < 80.5     to the left,  agree=0.704, adj=0.189, (0 split)
      ICM             < 24.05658 to the left,  agree=0.680, adj=0.123, (0 split)
      GGT             < 22.5     to the left,  agree=0.666, adj=0.086, (0 split)

Node number 6: 560 observations
  predicted class=0  expected loss=0.15  P(node) =0.1487779
    class counts:   476    84
   probabilities: 0.850 0.150

Node number 7: 973 observations,    complexity param=0.004074074
  predicted class=0  expected loss=0.2795478  P(node) =0.2585016
    class counts:   701   272
   probabilities: 0.720 0.280
  left son=14 (916 obs) right son=15 (57 obs)
  Primary splits:
      Trigliceridos   < 298.5    to the left,  improve=7.110452, (0 missing)
      RatioColesterol < 7.229064 to the left,  improve=5.146696, (0 missing)
      HDL.Colesterol  < 55.5     to the right, improve=4.990361, (0 missing)
      ICM             < 29.97304 to the left,  improve=4.181742, (0 missing)
      Glucosa         < 114.5    to the left,  improve=3.557733, (0 missing)
  Surrogate splits:
      RatioColesterol < 6.027027 to the left,  agree=0.950, adj=0.140, (0 split)
      HDL.Colesterol  < 32.5     to the right, agree=0.943, adj=0.035, (0 split)

Node number 14: 916 observations,    complexity param=0.004074074
  predicted class=0  expected loss=0.2652838  P(node) =0.2433581
    class counts:   673   243
   probabilities: 0.735 0.265
```

```
  left son=28 (833 obs) right son=29 (83 obs)
  Primary splits:
      Glucosa              < 114.5    to the left,  improve=3.800621, (0 missing)
      ICM                  < 29.97304 to the left,  improve=3.318296, (0 missing)
      HDL.Colesterol       < 55.5     to the right, improve=3.315045, (0 missing)
      ApolipoproteinaB     < 151.5    to the right, improve=3.051597, (0 missing)
      PulsacionesPorMinuto < 51.5     to the left,  improve=3.013770, (3 missing)
  Surrogate splits:
      MedicaDiabetes  splits as  LR,           agree=0.916, adj=0.072, (0 split)
      PresionSitolica < 171      to the left,  agree=0.913, adj=0.036, (0 split)

Node number 15: 57 observations,    complexity param=0.004074074
  predicted class=1  expected loss=0.4912281  P(node) =0.01514346
    class counts:    28    29
   probabilities: 0.491 0.509
  left son=30 (16 obs) right son=31 (41 obs)
  Primary splits:
      T4libre           < 0.705    to the left,  improve=4.854184, (0 missing)
      Creatinina        < 0.875    to the left,  improve=3.759851, (0 missing)
      ProteinaCReactiva < 0.075    to the right, improve=3.448921, (0 missing)
      PresionSitolica   < 113.5    to the right, improve=2.777790, (0 missing)
      ApolipoproteinaB  < 147      to the left,  improve=2.185137, (0 missing)
  Surrogate splits:
      GGT              < 32.5     to the left,  agree=0.807, adj=0.312, (0 split)
      TSH              < 2.655    to the right, agree=0.789, adj=0.250, (0 split)
      AcidoUrico       < 8.7      to the right, agree=0.772, adj=0.187, (0 split)
      BilirrubinaTotal < 0.35     to the left,  agree=0.754, adj=0.125, (0 split)
      HDL.Colesterol   < 55.5     to the right, agree=0.754, adj=0.125, (0 split)

Node number 28: 833 observations
  predicted class=0  expected loss=0.2521008  P(node) =0.2213071
    class counts:   623   210
   probabilities: 0.748 0.252

Node number 29: 83 observations,    complexity param=0.004074074
  predicted class=0  expected loss=0.3975904  P(node) =0.02205101
    class counts:    50    33
   probabilities: 0.602 0.398
  left son=58 (69 obs) right son=59 (14 obs)
  Primary splits:
      HDL.Colesterol    < 44.5     to the right, improve=5.310275, (0 missing)
      CreatininaEnOrina < 214.6    to the right, improve=3.483325, (8 missing)
      BilirrubinaTotal  < 0.46     to the right, improve=3.424758, (0 missing)
      GGT               < 54.5     to the left,  improve=2.749857, (0 missing)
      Edad              < 51.5     to the left,  improve=2.706700, (0 missing)
  Surrogate splits:
      RatioColesterol < 5.202564 to the left,  agree=0.880, adj=0.286, (0 split)
      Glucosa         < 155      to the left,  agree=0.855, adj=0.143, (0 split)
      Trigliceridos   < 209.5    to the left,  agree=0.855, adj=0.143, (0 split)

Node number 30: 16 observations
```

```
  predicted class=0  expected loss=0.1875  P(node) =0.004250797
    class counts:    13     3
   probabilities: 0.812 0.187

Node number 31: 41 observations
  predicted class=1  expected loss=0.3658537  P(node) =0.01089267
    class counts:    15    26
   probabilities: 0.366 0.634

Node number 58: 69 observations,    complexity param=0.004074074
  predicted class=0  expected loss=0.3188406  P(node) =0.01833156
    class counts:    47    22
   probabilities: 0.681 0.319
  left son=116 (58 obs) right son=117 (11 obs)
  Primary splits:
      GGT               < 54.5     to the left,  improve=2.829841, (0 missing)
      CreatininaEnOrina < 214.3    to the right, improve=2.697412, (6 missing)
      PresionSitolica   < 146      to the left,  improve=2.669560, (1 missing)
      BilirrubinaTotal  < 0.46     to the right, improve=2.432696, (0 missing)
      ProteinaCReactiva < 0.095    to the right, improve=2.302568, (0 missing)
  Surrogate splits:
      AST.GOT       < 40.5     to the left,  agree=0.870, adj=0.182, (0 split)
      Trigliceridos < 252.5    to the left,  agree=0.855, adj=0.091, (0 split)
      AcidoUrico    < 8.25     to the left,  agree=0.855, adj=0.091, (0 split)

Node number 59: 14 observations
  predicted class=1  expected loss=0.2142857  P(node) =0.003719447
    class counts:     3    11
   probabilities: 0.214 0.786

Node number 116: 58 observations
  predicted class=0  expected loss=0.2586207  P(node) =0.01540914
    class counts:    43    15
   probabilities: 0.741 0.259

Node number 117: 11 observations
  predicted class=1  expected loss=0.3636364  P(node) =0.002922423
    class counts:     4     7
   probabilities: 0.364 0.636


n= 3764

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 3764 675 0 (0.8206695 0.1793305)
    2) ProteinaCReactiva>=0.115 2231 319 0 (0.8570148 0.1429852) *
    3) ProteinaCReactiva< 0.115 1533 356 0 (0.7677756 0.2322244)
      6) ApolipoproteinaB< 91.95 560  84 0 (0.8500000 0.1500000) *
      7) ApolipoproteinaB>=91.95 973 272 0 (0.7204522 0.2795478)
       14) Trigliceridos< 298.5 916 243 0 (0.7347162 0.2652838)
```

```
        28) Glucosa< 114.5 833 210 0 (0.7478992 0.2521008) *
        29) Glucosa>=114.5 83  33 0 (0.6024096 0.3975904)
          58) HDL.Colesterol>=44.5 69  22 0 (0.6811594 0.3188406)
           116) GGT< 54.5 58  15 0 (0.7413793 0.2586207) *
           117) GGT>=54.5 11   4 1 (0.3636364 0.6363636) *
          59) HDL.Colesterol< 44.5 14   3 1 (0.2142857 0.7857143) *
       15) Trigliceridos>=298.5 57  28 1 (0.4912281 0.5087719)
        30) T4libre< 0.705 16   3 0 (0.8125000 0.1875000) *
        31) T4libre>=0.705 41  15 1 (0.3658537 0.6341463) *
```

Modelo con el criterio Gini

```
Call:
rpart(formula = CasosE4 ~ ., data = BaseModels[, -34], method = "class",
    parms = list(split = "Gini"), control = rpart.control(xval = 10,
        cp = 0))
  n= 3764

             CP nsplit rel error    xerror       xstd
1  0.0051851852      0 1.0000000 1.0000000 0.03486842
2  0.0044444444     15 0.9140741 0.9985185 0.03484822
3  0.0038518519     16 0.9096296 1.0148148 0.03506885
4  0.0037037037     22 0.8859259 1.0829630 0.03595558
5  0.0029629630     26 0.8666667 1.0992593 0.03615935
6  0.0022222222     43 0.8103704 1.1614815 0.03690945
7  0.0019047619     64 0.7540741 1.2281481 0.03766628
8  0.0014814815     82 0.7155556 1.2518519 0.03792423
9  0.0012345679     89 0.7037037 1.3244444 0.03867952
10 0.0007407407     96 0.6918519 1.3585185 0.03901664
11 0.0004938272    114 0.6785185 1.3851852 0.03927297
12 0.0001139601    117 0.6770370 1.4162963 0.03956388
13 0.0000000000    130 0.6755556 1.4162963 0.03956388

Variable importance
        PesoTrabajador        RatioColesterol     PerimetroAbdominal
             51.492939              51.207659              46.313150
                   ICM              ApoBApoA1                ALT.GPT
             41.120181              40.766705              37.964054
         Trigliceridos                AST.GOT                    ICA
             37.861458              34.828191              31.662039
            Creatinina             Colesterol                Glucosa
             30.017042              28.760551              28.677132
         Microalbumina                T4libre      ProteinaCReactiva
             28.561288              27.888611              27.747288
      ApolipoproteinaB      PresionDiastolica        PresionSitolica
             27.163687              24.605592              23.678291
     CreatininaEnOrina                    GGT            AcidoUrico
             21.613641              18.998928              17.422763
                  Edad                    TSH         HDL.Colesterol
             15.937235              15.583868              15.411045
     ApolipoproteinaA1         Lipoproteina.a       BilirrubinaTotal
             13.639143              12.706117              11.483187
```

```
               Calcio HemoglobinaGlicosilada   PulsacionesPorMinuto
            10.485818               9.434361               6.433605
       ConsumoAlcohol         MedicaDiabetes      metSindrom.Recode
             5.791674               2.827120               1.342954
     MedicaDislipemia    MedicaAntiagregante
             1.307051               0.161225
```

Después de podar

```
 Call:
rpart(formula = CasosE4 ~ ., data = BaseModels[, -c(remove, 34)],
    method = "class", parms = list(split = "Gini"), control = rpart.control(xval = 10,
        cp = 0))
  n= 3764

          CP nsplit rel error    xerror       xstd
1 0.005185185      0 1.0000000 1.0000000 0.03486842
2 0.003851900      6 0.9659259 0.9985185 0.03484822

Variable importance
          ApoBApoA1    ProteinaCReactiva      RatioColesterol
          20.8852199           11.4522920            9.5372148
       Trigliceridos              Glucosa           Creatinina
           7.9547391            5.9048588            3.3719064
             ALT.GPT               Calcio       HDL.Colesterol
           2.0231438            1.7126485            1.7083776
   PresionDiastolica PulsacionesPorMinuto    metSindrom.Recode
           1.3487625            1.3487625            0.4188280
                 TSH           AcidoUrico                  GGT
           0.3781068            0.2767599            0.2135626
             T4libre     BilirrubinaTotal
           0.1890534            0.1868672

Node number 1: 3764 observations,    complexity param=0.005185185
  predicted class=0  expected loss=0.1793305  P(node) =1
    class counts:  3089   675
   probabilities: 0.821 0.179
  left son=2 (2438 obs) right son=3 (1326 obs)
  Primary splits:
      ApoBApoA1         < 0.8168477 to the left,  improve=14.559790, (14 missing)
      ProteinaCReactiva < 0.115     to the right, improve=14.492290, (2 missing)
      RatioColesterol   < 4.427249  to the left,  improve= 7.026025, (0 missing)
      Trigliceridos     < 280.5     to the left,  improve= 4.224513, (0 missing)
      HDL.Colesterol    < 42.5      to the right, improve= 4.208204, (0 missing)
  Surrogate splits:
      RatioColesterol   < 4.406406  to the left,  agree=0.869, adj=0.629, (14 split)
      Trigliceridos     < 146.5     to the left,  agree=0.701, adj=0.151, (0 split)
      HDL.Colesterol    < 43.5      to the right, agree=0.689, adj=0.117, (0 split)
      metSindrom.Recode splits as  LR, agree=0.658, adj=0.029, (0 split)
      AcidoUrico        < 7.85      to the left,  agree=0.651, adj=0.010, (0 split)

Node number 2: 2438 observations
  predicted class=0  expected loss=0.1472518  P(node) =0.6477152
```

```
    class counts:  2079   359
   probabilities: 0.853 0.147

Node number 3: 1326 observations,    complexity param=0.005185185
  predicted class=0  expected loss=0.2383107  P(node) =0.3522848
    class counts:  1010   316
   probabilities: 0.762 0.238
  left son=6 (897 obs) right son=7 (429 obs)
  Primary splits:
      ProteinaCReactiva    < 0.115     to the right, improve=11.452290, (0 missing)
      ApoBApoA1            < 0.8196601 to the right, improve= 3.949875, (5 missing)
      GGT                  < 28.5      to the right, improve= 3.743740, (0 missing)
      PulsacionesPorMinuto < 67.5      to the right, improve= 3.620959, (9 missing)
      Trigliceridos        < 126.5     to the right, improve= 3.144619, (0 missing)
  Surrogate splits:
      GGT              < 21.5      to the right, agree=0.683, adj=0.019, (0 split)
      BilirrubinaTotal < 1.62      to the left,  agree=0.682, adj=0.016, (0 split)
      AcidoUrico       < 2.65      to the right, agree=0.680, adj=0.012, (0 split)
      Trigliceridos    < 42        to the right, agree=0.678, adj=0.005, (0 split)
      Calcio           < 10.25     to the left,  agree=0.677, adj=0.002, (0 split)

Node number 6: 897 observations
  predicted class=0  expected loss=0.1928651  P(node) =0.2383103
    class counts:   724   173
   probabilities: 0.807 0.193

Node number 7: 429 observations,    complexity param=0.005185185
  predicted class=0  expected loss=0.3333333  P(node) =0.1139745
    class counts:   286   143
   probabilities: 0.667 0.333
  left son=14 (422 obs) right son=15 (7 obs)
  Primary splits:
      ApoBApoA1      < 0.8220618 to the right, improve=6.325434, (0 missing)
      Glucosa        < 131.5     to the left,  improve=5.674603, (0 missing)
      PesoTrabajador < 81.05     to the left,  improve=4.623503, (0 missing)
      T4libre        < 0.705     to the left,  improve=4.333333, (0 missing)
      Trigliceridos  < 374       to the left,  improve=4.325309, (0 missing)

Node number 14: 422 observations,    complexity param=0.005185185
  predicted class=0  expected loss=0.3222749  P(node) =0.1121148
    class counts:   286   136
   probabilities: 0.678 0.322
  left son=28 (413 obs) right son=29 (9 obs)
  Primary splits:
      Glucosa         < 131.5     to the left,  improve=5.904859, (0 missing)
      Trigliceridos   < 374       to the left,  improve=4.664097, (0 missing)
      PesoTrabajador  < 81.05     to the left,  improve=4.364213, (0 missing)
      T4libre         < 0.705     to the left,  improve=4.359332, (0 missing)
      RatioColesterol < 7.229064  to the left,  improve=4.072730, (0 missing)

Node number 15: 7 observations
```

```
  predicted class=1  expected loss=0  P(node) =0.001859724
    class counts:     0     7
   probabilities: 0.000 1.000

Node number 28: 413 observations,    complexity param=0.005185185
  predicted class=0  expected loss=0.3099274  P(node) =0.1097237
    class counts:   285   128
   probabilities: 0.690 0.310
  left son=56 (390 obs) right son=57 (23 obs)
  Primary splits:
      Trigliceridos   < 374       to the left,  improve=4.348228, (0 missing)
      T4libre         < 0.705     to the left,  improve=4.341708, (0 missing)
      RatioColesterol < 7.229064  to the left,  improve=4.264507, (0 missing)
      PesoTrabajador  < 81.05     to the left,  improve=4.003666, (0 missing)
      HDL.Colesterol  < 63.5      to the right, improve=3.213959, (0 missing)
  Surrogate splits:
      TSH             < 6.405     to the left,  agree=0.949, adj=0.087, (0 split)
      RatioColesterol < 7.597403  to the left,  agree=0.949, adj=0.087, (0 split)
      T4libre         < 0.595     to the right, agree=0.947, adj=0.043, (0 split)

Node number 29: 9 observations
  predicted class=1  expected loss=0.1111111  P(node) =0.002391073
    class counts:     1     8
   probabilities: 0.111 0.889

Node number 56: 390 observations
  predicted class=0  expected loss=0.2923077  P(node) =0.1036132
    class counts:   276   114
   probabilities: 0.708 0.292

Node number 57: 23 observations,    complexity param=0.005185185
  predicted class=1  expected loss=0.3913043  P(node) =0.006110521
    class counts:     9    14
   probabilities: 0.391 0.609
  left son=114 (10 obs) right son=115 (13 obs)
  Primary splits:
      Creatinina        < 0.985     to the left,  improve=3.371906, (0 missing)
      CreatininaEnOrina < 160.4     to the left,  improve=3.371906, (0 missing)
      ApoBApoA1         < 1.087261  to the left,  improve=3.081522, (0 missing)
      AcidoUrico        < 6.95      to the right, improve=2.242236, (0 missing)
      GGT               < 51        to the right, improve=1.339855, (0 missing)
  Surrogate splits:
      ALT.GPT              < 38        to the right, agree=0.826, adj=0.6, (0 split)
      Calcio               < 9.15      to the left,  agree=0.783, adj=0.5, (0 split)
      Trigliceridos        < 618.5     to the right, agree=0.739, adj=0.4, (0 split)
      PresionDiastolica    < 88.5      to the right, agree=0.739, adj=0.4, (0 split)
      PulsacionesPorMinuto < 77        to the right, agree=0.739, adj=0.4, (0 split)

Node number 114: 10 observations
  predicted class=0  expected loss=0.3  P(node) =0.002656748
    class counts:     7     3
```

```
   probabilities: 0.700 0.300

Node number 115: 13 observations
  predicted class=1  expected loss=0.1538462  P(node) =0.003453773
    class counts:     2    11
   probabilities: 0.154 0.846
```

```
n= 3764

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 3764 675 0 (0.8206695 0.1793305)
    2) ApoBApoA1< 0.8168477 2438 359 0 (0.8527482 0.1472518) *
    3) ApoBApoA1>=0.8168477 1326 316 0 (0.7616893 0.2383107)
      6) ProteinaCReactiva>=0.115 897 173 0 (0.8071349 0.1928651) *
      7) ProteinaCReactiva< 0.115 429 143 0 (0.6666667 0.3333333)
       14) ApoBApoA1>=0.8220618 422 136 0 (0.6777251 0.3222749)
         28) Glucosa< 131.5 413 128 0 (0.6900726 0.3099274)
           56) Trigliceridos< 374 390 114 0 (0.7076923 0.2923077) *
           57) Trigliceridos>=374 23   9 1 (0.3913043 0.6086957)
            114) Creatinina< 0.985 10   3 0 (0.7000000 0.3000000) *
            115) Creatinina>=0.985 13   2 1 (0.1538462 0.8461538) *
         29) Glucosa>=131.5 9   1 1 (0.1111111 0.8888889) *
       15) ApoBApoA1< 0.8220618 7   0 1 (0.0000000 1.0000000) *
```

B.3. Regresión penalizada

Coeficientes con todas las variables

Modelo	(Intercept)	Lipoproteina.a	ProteinaCReactiva	ApoBApoA1	RatioColesterol	Glucosa.MDis
CV.Lasso	-2.11E+00	7.44E-04	-2.02E-01	7.92E-01	1.11E-02	4.72E-05
CV.net.08	-2.10E+00	7.34E-04	-1.86E-01	7.72E-01	1.18E-02	7.09E-05
Modelo	**Lipoproteina.a.MAnt**	**ConsumoAlcohol.MDis**	**GlucosaEF[T.Alto]**	**ProteinaCReactivaEF[T.Alto]**	**TrigliceridosEF[T.Normal]**	
CV.Lasso	9.48E-04	6.02E-04	1.02E-02	-1.18E-01	-4.25E-02	
CV.net.08	9.19E-04	5.61E-04	9.89E-03	-1.24E-01	-4.03E-02	

Tabla B.4: Coeficientes de la regresión penalizada para los modelos Lasso y .net con un coeficiente alfa de 0.8

CV.net.06					
(Intercept)	ApolipoproteinaB	Lipoproteina.a	ProteinaCReactiva	ApoBApoA1	RatioColesterol
-2.1011	0.0004	0.0007	-0.1673	0.7099	0.0126
Glucosa.MDis	Lipoproteina.a.MAnt	ConsumoAlcohol.MDis	GlucosaEF[T.Alto]	ProteinaCReactivaEF[T.Alto]	TrigliceridosEF[T.Normal]
0.0001	0.0009	0.0005	0.0086	-0.1289	-0.0374

CV.net.04					
(Intercept)	ApolipoproteinaA1	ApolipoproteinaB	Lipoproteina.a	ProteinaCReactiva	ApoBApoA1
-2.0804	-0.0005	0.0012	0.0009	-0.2026	0.6364
RatioColesterol	Glucosa.MDis	Lipoproteina.a.MAnt	Microalbumina.metSin	ConsumoAlcohol.MDis	GlucosaEF[T.Alto]
0.0158	0.0002	0.0019	0.0002	0.0007	0.0292
Lipoproteina.aEF[T.Alto]	ProteinaCReactivaEF[T.Alto]	PresionSitolicaEF[T.Alto]			
0.0153	-0.1476	0.0262			

CV.net.02					
(Intercept)	ApolipoproteinaA1	ApolipoproteinaB	Lipoproteina.a	ProteinaCReactiva	TSH
-1.9461	-0.0014	0.0021	0.0008	-0.1702	-0.0015
ApoBApoA1	RatioColesterol	Glucosa.MDis	Lipoproteina.a.MAnt	Microalbumina.metSin	ConsumoAlcohol.MDis
0.4759	0.0173	0.0003	0.0021	0.0003	0.0007
BilirrubinaTotalEF[T.Alto]	GlucosaEF[T.Alto]	Lipoproteina.aEF[T.Alto]	ProteinaCReactivaEF[T.Alto]	PresionSitolicaEF[T.Alto]	TrigliceridosEF[T.Normal]
0.0086	0.0337	0.0263	-0.1589	0.0373	-0.0801

Tabla B.5: Coeficientes de la regresión penalizada para los modelos .net con coeficientes alfa de 0.6, 0.4 y 0.2 respectivamente.

Coeficientes con las variables de interación

CV.lasso.SEF					
(Intercept)	Lipoproteina.a	ProteinaCReactiva	ApoBApoA1	RatioColesterol	Glucosa.MDis
-2.1439	0.0012	-0.3917	0.8323	0.0135	0.0002
Lipoproteina.a.MAnt	Microalbumina.metSin	ConsumoAlcohol.MDis			
0.0021	0.0002	0.0009			

CV.net.08.SEF					
(Intercept)	Lipoproteina.a	ProteinaCReactiva	ApoBApoA1	RatioColesterol	Glucosa.MDis
-2.1366	0.0011	-0.3710	0.8137	0.0140	0.0002
Lipoproteina.a.MAnt	Microalbumina.metSin	ConsumoAlcohol.MDis			
0.0020	0.0002	0.0008			

CV.net.06.SEF					
(Intercept)	ApolipoproteinaB	Lipoproteina.a	ProteinaCReactiva	ApoBApoA1	RatioColesterol
-2.1383	0.0004	0.0011	-0.3413	0.7473	0.0147
Glucosa.MDis	Lipoproteina.a.MAnt	Microalbumina.metSin	ConsumoAlcohol.MDis		
0.0003	0.0020	0.0002	0.0008		

CV.net.04.SEF					
(Intercept)	ApolipoproteinaA1	ApolipoproteinaB	Lipoproteina.a	ProteinaCReactiva	ApoBApoA1
-2.0898	-0.0003	0.0011	0.0010	-0.2962	0.6256
RatioColesterol	Glucosa.MDis	Lipoproteina.a.MAnt	Microalbumina.metSin	ConsumoAlcohol.MDis	
0.0154	0.0003	0.0019	0.0002	0.0007	

CV.net.02.SEF					
(Intercept)	ApolipoproteinaA1	ApolipoproteinaB	Lipoproteina.a	ProteinaCReactiva	TSH
-1.9589	-0.0012	0.0020	0.0011	-0.2457	-0.0022
ApoBApoA1	RatioColesterol	Glucosa.MDis	Lipoproteina.a.MAnt	Microalbumina.metSin	ConsumoAlcohol.MDis
0.4647	0.0168	0.0004	0.0022	0.0003	0.0007

Tabla B.6: Coeficientes de la regresión penalizada para los modelos Lasso y .net con coeficientes alfa de 0.8, 0.6, 0.4 y 0.2 respectivamente.

Coeficientes con las variables originales

CV.lasso.SEF					
(Intercept)	Lipoproteina.a	ProteinaCReactiva	MedicaDislipemia	ApoBApoA1	RatioColesterol
-2.1445	0.0013	-0.3870	0.0677	0.8235	0.0155

CV.net.08.SEF						
(Intercept)	ApolipoproteinaB	Lipoproteina.a	ProteinaCReactiva	MedicaDislipemia	ApoBApoA1	RatioColesterol
-2.1390	0.0001	0.0012	-0.3666	0.0658	0.7997	0.0160

CV.net.06.SEF						
(Intercept)	ApolipoproteinaB	Lipoproteina.a	ProteinaCReactiva	MedicaDislipemia	ApoBApoA1	RatioColesterol
-2.1406	0.0005	0.0012	-0.3374	0.0622	0.7344	0.0164

CV.net.04.SEF						
(Intercept)	ApolipoproteinaA1	ApolipoproteinaB	Lipoproteina.a	ProteinaCReactiva	TSH	MedicaDislipemia
-2.0657	-0.0006	0.0014	0.0013	-0.3292	-0.0027	0.0757
ApoBApoA1	RatioColesterol					
0.6159	0.0179					

CV.net.02.SEF						
(Intercept)	ApolipoproteinaA1	ApolipoproteinaB	Lipoproteina.a	ProteinaCReactiva	TSH	MedicaDislipemia
-1.9485	-0.0014	0.0022	0.0013	-0.2721	-0.0050	0.0805
ApoBApoA1	RatioColesterol					
0.4677	0.0191					

Tabla B.7: Coeficientes de la regresión penalizada para los modelos Lasso y .net con coeficientes alfa de 0.8, 0.6, 0.4 y 0.2 respectivamente.

B.4. Modelos de particionamiento recursivo

Primer modelo

```
Model-based recursive partitioning (logit)

Model formula:
CasosE4 ~ ApoBApoA1 + ProteinaCReactiva + Lipoproteina.a | RatioColesterol +
    CreatininaEnOrina + Calcio + ApolipoproteinaA1 + BilirrubinaTotal +
    Colesterol + HDL.Colesterol + Creatinina + GGT + Glucosa +
    AST.GOT + ALT.GPT + Trigliceridos + AcidoUrico + ApolipoproteinaB +
    T4libre + TSH + HemoglobinaGlicosilada + Microalbumina +
    PesoTrabajador + PerimetroAbdominal + PresionSitolica + PresionDiastolica +
    PulsacionesPorMinuto + ConsumoAlcohol + Edad + ICM +
    ICA + MedicaDiabetes + MedicaDislipemia + MedicaHipertension +
    MedicaAntiagregante + metSindrom.Recode

Fitted party:
[1] root: n = 2833
           (Intercept)       x(Intercept)         xApoBApoA1
           -2.34395792                 NA         1.27818339
    xProteinaCReactiva    xLipoproteina.a
           -0.95543782         0.00342053

Number of inner nodes:    0
Number of terminal nodes: 1
Number of parameters per node: 5
Objective function: 1308.125

-- Node 1 ---------------------------------
Number of observations: 2833
```

```
Parameter instability tests:
          RatioColesterol CreatininaEnOrina     Calcio BilirrubinaTotal
statistic        10.32388          10.09716 10.7324986       11.0339247
p.value           1.00000           1.00000  0.9999996        0.9999985
          Colesterol HDL.Colesterol Creatinina      GGT  Glucosa    AST.GOT
statistic 10.7423534      10.941287   8.168835 7.801774 9.604353 12.0350749
p.value    0.9999996       0.999999   1.000000 1.000000 1.000000  0.9999236
           ALT.GPT Trigliceridos AcidoUrico ApolipoproteinaA1 ApolipoproteinaB
statistic 7.263587    15.9147172   8.169273        10.8103440         10.04334
p.value   1.000000     0.9032777   1.000000         0.9999995          1.00000
             T4libre      TSH HemoglobinaGlicosilada Microalbumina
statistic 13.0566464 4.390599               9.219104        10.221
p.value    0.9986742 1.000000               1.000000         1.000
          PesoTrabajador PerimetroAbdominal PresionSitolica PresionDiastolica
statistic       8.026393         10.6995341        8.938464          6.687576
p.value         1.000000          0.9999997        1.000000          1.000000
          PulsacionesPorMinuto ConsumoAlcohol       Edad      ICM      ICA
statistic           13.9585935       9.143921 12.5029557 8.982143 7.713097
p.value              0.9917717       1.000000  0.9996817 1.000000 1.000000
          MedicaDiabetes MedicaDislipemia MedicaHipertension
statistic       2.744975        5.8292777           1.497464
p.value         1.000000        0.9996193           1.000000
          MedicaAntiagregante metSindrom.Recode
statistic           5.6210319         5.7022647
p.value             0.9998149         0.9997529

Best splitting variable: Trigliceridos
Perform split? no


Call:
glm(formula = y ~ x, family = binomial, start = start)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7462  -0.6601  -0.5926  -0.4970   2.6786

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.343958   0.188308 -12.447  < 2e-16 ***
xApoBApoA1          1.278183   0.229132   5.578 2.43e-08 ***
xProteinaCReactiva -0.955438   0.211151  -4.525 6.04e-06 ***
xLipoproteina.a     0.003421   0.001554   2.201   0.0278 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2674.1  on 2832  degrees of freedom
Residual deviance: 2616.2  on 2829  degrees of freedom
```

```
AIC: 2624.2

Number of Fisher Scoring iterations: 5
```

Segundo modelo

```
Model-based recursive partitioning (logit)

Model formula:
CasosE4 ~ ApoBApoA1 + ProteinaCReactiva + Lipoproteina.a + RatioColesterol |
    ApolipoproteinaB + GGT + ApolipoproteinaA1 + PresionSitolica +
        ICM + HemoglobinaGlicosilada

Fitted party:
[1] root
|   [2] ApolipoproteinaB <= 121: n = 2410
|              (Intercept)       x(Intercept)         xApoBApoA1
|             -2.683262484                 NA        3.483606648
|          xLipoproteina.a   xRatioColesterol xProteinaCReactiva
|              0.003097758       -0.312086274       -0.804152431
|   [3] ApolipoproteinaB > 121: n = 856
|              (Intercept)       x(Intercept)         xApoBApoA1
|            -1.7532282427                 NA       0.1271806642
|          xLipoproteina.a   xRatioColesterol xProteinaCReactiva
|            -0.0006306751       0.0885454557      -0.3546838101

Number of inner nodes:    1
Number of terminal nodes: 2
Number of parameters per node: 6
Objective function: 1497.855

-- Node 1 ---------------------------------
Number of observations: 3266

Parameter instability tests:
          ApolipoproteinaB       GGT ApolipoproteinaA1
statistic      26.49355217 9.9238895         9.4452264
p.value         0.01296412 0.9971177         0.9988784

        PresionSitolica        ICM HemoglobinaGlicosilada
statistic    12.6360394 12.2426526              9.7912312
p.value       0.9160395  0.9399255              0.9977499

Best splitting variable: ApolipoproteinaB
Perform split? yes
Selected split: <= 121 | > 121

-- Node 2 ---------------------------------
Number of observations: 2410
```

```
Parameter instability tests:
          ApolipoproteinaB        GGT ApolipoproteinaA1
statistic       10.6161354 11.1501799         17.655393
p.value          0.9910122  0.9813709          0.358456

           PresionSitolica       ICM HemoglobinaGlicosilada
statistic       15.9740904 8.4244476             12.1301283
p.value          0.5543907 0.9999141              0.9459067
Best splitting variable: ApolipoproteinaA1
Perform split? no

-- Node 3 ---------------------------------
Number of observations: 856

Parameter instability tests:
          ApolipoproteinaB        GGT ApolipoproteinaA1
statistic       10.1656984 10.9672712         6.9330629
p.value          0.9955517  0.9851688         0.9999997

            PresionSitolica        ICM HemoglobinaGlicosilada
statistic       15.7098266 13.3138842              7.1428870
p.value          0.5870096  0.8624425              0.9999993

Best splitting variable: PresionSitolica
Perform split? no
```

Modelos intermedios del modelo 2

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-2.683	0.290	-9.241	<2E-16	***
xApoBApoA1	3.484	0.661	5.269	0.000	***
xProteinaCReactiva	-0.804	0.219	-3.676	0.000	***
xLipoproteina.a	0.003	0.002	1.703	0.089	.
xRatioColesterol	-0.312	0.125	-2.500	0.012	*

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-2.956	0.268	-11.046	<2E-16	***
ApoBApoA1	2.116	0.371	5.701	0.000	***
Lipoproteina.a	0.004	0.002	2.036	0.042	*
ProteinaCReactiva	-0.832	0.222	-3.753	0.000	***

Tabla B.8: Secuencias de tablas para el refinamiento del modelo 2 del particionamiento recursivo basado en modelos.

Modelos intermedios del modelo 3

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-1.753	0.504	-3.482	0.000	***
xApoBApoA1	0.127	0.560	0.227	0.820	
xProteinaCReactiva	-0.355	0.295	-1.203	0.229	
xLipoproteina.a	-0.001	0.002	-0.257	0.797	
xRatioColesterol	0.089	0.089	0.992	0.321	

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-1.686	0.413	-4.083	0.000	***
RatioColesterol	0.100	0.078	1.289	0.197	
Lipoproteina.a	-0.001	0.002	-0.259	0.796	
ProteinaCReactiva	-0.353	0.295	-1.198	0.231	

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	-1.710	0.403	-4.241	0.000	***
RatioColesterol	0.101	0.078	1.302	0.193	
ProteinaCReactiva	-0.356	0.295	-1.204	0.228	

Coeficiente	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-3.044	2.388	-1.275	0.202
RatioColesterol	0.253	0.393	0.643	0.520

Tabla B.9: Secuencias de tablas para el refinamiento del modelo 3 del particionamiento recursivo basado en modelos.